

# Signal Process Classification in a High Dimensional Setting

Thesis submitted for the degree of “Doctor of Philosophy”

by

**Jonathan Sidi**

Submitted to the Senate of the Hebrew University

May 2018

This work was carried out under the supervision of  
Professor Ya'acov Ritov

## ABSTRACT

Informed policy decision-making in the age of high-dimensional data has become increasingly difficult as data storage capabilities have grown in the last two decades. A causality dilemma presents itself in which policy questions necessitate methodological improvements in statistical inference and prediction, but on the other hand improved methodology and prediction opens new avenues of inquiry which drive new policy. This thesis examines how improvements of signal extraction in complex high-dimensional data can affect the inference and decision-making in a time sensitive policy environment. This thesis consists of three main chapters which explore various applied, technical and theoretical issues in statistics which are derived to solve concrete policy problems.

Following the introduction, which provides context for each chapter, Chapter 2 deals with the problem of ‘nowcasting’ which attempts to predict the present. This may seem counter-intuitive at first since we are always in the present, but many policy decisions hinge of the timeliness of data availability at the time a decision is made. Regardless of availability of information a predetermined deadline forces a decision to be made. There are a number of fields that must determine the current state of a system while still accumulating information such as economics, epidemiology, finance, meteorology and social networks. As technology has improved the ability to measure and retain larger amounts of data, the definition of the present, i.e. ‘now’, has been refined to greater granularity. This problem is more pronounced when the system being predicted is a projection based on high dimensional data, where the number of predictors is larger than the number of observations. In this case, least squares is not feasible and dimension reduction must take place. We survey a number of methods in this chapter to reduce the dimension of data comparing two major schools of practice - dimension reduction conditional and unconditional on the predicted variable. Throughout the chapter the theoretical comparison is complemented by the policy challenge of setting the interest rate of the Israeli economy by the central Bank of Israel.

Chapter 3 deals with another facet of the continuous accumulation of data. We look at data that is itself continuously revised, thereby creating a setting where

---

there is no final value of the target process. This creates a new layer of uncertainty to take into account when arriving at a position, how much will the current data point and thereby the entire series be different within the upcoming periods of time. This can have serious ramifications when a change in policy is undertaken. Such a scenario would be presented with a certain reality regarding the state of a process and reacting with proper steps, while in retrospect the data representing the same time points in the series was subsequently revised to show a different reality only periods of time later, rendering the policy decision inaccurate, possibly resulting in a negative impact. In this chapter, we propose a method to estimate the uncertainty in the revision process. We use this estimation to generate an asymmetrical prediction interval of a subsequent revision to the currently published activity period. The interval is a function of the maturity of each time period within the current vintage of the growth process. Our approach relaxes a common assumption in the literature that the revision process is a homogeneous one and not a mixture of a number of different processes. We postulate that the revision process is a function of the state of the growth process, thereby creating the necessity to model the process state through hidden Markov models and estimate model parameters relating to each one. The methodology is tested on historical data of the Gross Domestic Product (GDP) of the Israel. We find that there is a significant difference in the size and sign of revisions dependent on the state of the growth rate. More specifically, when the initial publication is in a low growth period, the growth is overestimated and subsequent revisions lower the growth rate, and conversely when the initial publication is in a high growth period the growth is underestimated and the subsequent revisions increase the growth rate.

Chapter 4 discusses regularization and classification of linear mixed models. Such data structures were originally beyond the scope of the initial research that produced generalized linear model regularization such as the  $\ell_1$  and  $\ell_2$  families of model selection. Regularization of the mixed effects models allows researchers the flexibility to model more complex data structures with subject specific random effects. These types of data structures are prevalent in applied fields of studies and allow capturing the complexity observed in the real world. This chapter will define an extension to the current set of penalties researched in linear mixed models and generalized linear mixed models, the Linear Mixed Model Elastic Net (LMMEN) penalty. The goal of the penalty is to simultaneously select both the fixed and random effects in the model while allowing for high levels of correlation among the either type of effect. Theoretical results and simulations comparing various

competing penalties are discussed in the chapter. Efficiently and accurately aggregating crowd sentiment to answer probabilistic questions that shape policy is a challenge that was introduced by the lowering of the cost of having a direct and continuous channel to a larger group of people. Harnessing the knowledge of such a population with varying characteristics and aggregating it into a single forecast was the subject of the case study we test the LMMEN on. This randomized control study accumulated longitudinal data where probabilistic forecasts are derived from crowd sentiment to answer various economic and geopolitical questions of interest.

Taken together, the chapters in this work span different aspects of how statistical challenges found in complex high dimensional data has helped shape the new direction of informed policy-decision making. All chapters have been submitted for publication at peer reviewed journals and are currently under review. Chapter 2 has been published as a discussion paper in the Bank of Israel and both the methodology presented in it and in chapter 3 are currently applied in the Bank of Israel as part of the monthly assessment of the state of the Israeli economy used to derive the decision of the central bank's interest rates adjustments.

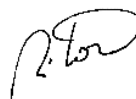
## A LETTER OF CONTRIBUTION

### *“Signal Process Classification in a High Dimensional Setting”*

**Author:** Jonathan Sidi

**Advisor:** Prof. Ya’acov Ritov

**Signature:**



### ***Chapter 2: “Nowcasting Israel GDP Using High Frequency Macroeconomic Disaggregates”***

Coauthor of this chapter is Gil Dafnai from the Economics department at the Hebrew University. J.S., together with G.D., designed the research. The statistical subject matter, simulation study and data analysis were conducted by J.S, under guidance of Y.R. The economic subject matter in the case studies were designed by G.D. and implemented by J.S.

### ***Chapter 3: “Estimating the Uncertainty of Data Revisions in Asynchronous Stationary Processes using Hidden Markov Models”***

Chapter 3 was written under the guidance of Y.R. Development of research topic, estimation methods, simulations and the writing of the paper was carried out by J.S. under guidance of Y.R.

### ***Chapter 4: “Regularization and Classification of Linear Mixed Models via the Elastic Net Penalty with Application to the Good Judgment Project”***

Coauthors of this chapter are Y.R. and Lyle Ungar of the Computer and Information Statistics department of the University of Pennsylvania. Y.R. designed the initial estimation methods and L.U. supplied the data source. The development of the extended methodology, simulation and writing of the paper was carried out by J.S. under the guidance of Y.R.. The development and publication of the accompanying R package for the paper was authored and submitted to CRAN by J.S..

## CONTENTS

1. <i>Introduction</i> . . . . .	8
2. <i>Nowcasting Israel GDP Using High Frequency Macroeconomic Disaggregates</i> . . . . .	22
3. <i>Estimating the Uncertainty of Data Revisions in Asynchronous Stationary Processes using Hidden Markov Models</i> . . . . .	79
4. <i>Regularization and Classification of Linear Mixed Models via the Elastic Net Penalty with Application to the Good Judgment Project</i> . . . . .	111
5. <i>Discussion</i> . . . . .	160

## 1. INTRODUCTION



Rapid advancements in the ability to acquire and store vast amount of data in the last two decades has shifted the research of high dimensional data from the exception to the standard. To accommodate this change a cross pollination of various fields of study has occurred paving new research avenues, and informed policy decision making is no exception to this phenomenon.

The symbiotic relationship between leading public policy institutions and high dimensional methodology has spread from academia to include government agencies and the private sector at an increasing pace along with the ability to store and investigate granular data in the last decade. Examples of such institutions are the Computation Institute and Harris School of Public Policy at the University of Chicago, the Institute for Social Research at University of Michigan, the eScience Institute of the University of Washington, and most recently, the Massive Data Institute at Georgetown University. A common thread that runs through these institutions is the “Data Science for Social Good” initiative which partners data scientists from quantitative fields with nonprofits and government agencies, to tackle data-intensive high impact problems in education, public health, public safety, criminal justice, environmental issues, city operations, and social services, University of Chicago (2017).

Government agencies, which have a direct line to derive and enact public policy, have followed suit. Investments in technological infrastructure have allowed agencies such as central banks, law enforcement, health and human services and statistical agencies to retain, analyze and derive policy through the performance based program assessments.

This dissertation examines how improvements in signal extraction for complex high-dimensional data can affect inference and decision-making in a time sensitive policy environment. The research explores various applied,

technical and theoretical issues in statistics which are derived to solve real-world policy problems. Following this introductory chapter, chapters two and three focus on high dimension methodological problems encountered within central banks in the process of deriving public policy to regulate a national economy. Chapter 4 focuses on how a time sensitive policy problem necessitated the development of a new regularization penalty for linear mixed model in a high dimensional setting and we conclude with discussions.

## 1 Methodology Driving Policy

The main lever which a central bank has at its disposal in regulation is the benchmark interest rate. The interest rate is set every month on a fixed day in order to create a consistent expectation that the markets can depend on. In order to derive a decision, data are gathered from a wide array of sources: labor surveys, financial sectors, manufacturing sectors, price indices, and the global markets. Each source has hundreds of variables to track with mixed frequencies and varying series length. Most importantly series are updated at different dates and the agencies collecting the data may only release information relating to the date of activity many time periods after it has passed, creating a time lag between the process and the publication date. These sources are continuously collected, analyzed and aggregated to create a picture of what the current state of the economy by various departments and summarized for the banks' board to derive a decision on what policy to enact.

We narrow our scope to an aggregate, the gross domestic product (GDP), which represents the total output of the economy. This series is a nominal series (i.e. units are denoted as monetary) and is updated once every three

months or, once a quarter. The percent change in the series is used to indicate the relative growth of the economy, which is interpreted as its health. At time of publication the marginal datum relates to the total output five weeks prior, and is revised thereafter every month for approximately five years. This in essence makes the entire history of the GDP a random variable and every month the published series is a realization of the data generating process.

These revisions include new information relevant to the activity period of initial publication that were not available until the current period. Revisions may be separated into three major groups: soft data sources, hard data sources and methodological improvements of measurements. Soft data include surveys of the labor, business and trade sectors that are acquired over long periods of time and relevant information from them could be derived anywhere from a month to a few years after the initial publication period has passed. Hard data are actual measurements that different agencies procure, such as tax revenue, government expenditures and private consumption. Methodological improvements include the methods used in surveys, the manner in which data is collected and improvements in statistical methodology such as seasonal adjustment of data.

This difficulty is compounded by the bank's hard deadline to update their interest rate. All analysis converges to a single meeting and the timeliness of data is paramount. Any series with updated data at hand is scrutinized and the performance of series yet to published are estimated. The importance of the GDP as a source to estimate the growth of the economy creates a demand to have a current estimate of it at every meeting which the interest rate is set. We focus on two areas of weaknesses attributed to the GDP, the five week lag inherent in the initial publication of a new data point of the series, chapter 2, and the estimation of the uncertainty found in the GDP

growth process created by its continuous revision, chapter 3.

Chapter 2 deals with the problem of prediction when there are more variables than observations. Consider a linear regression model

$$Y = X\beta + \epsilon \quad (1)$$

where  $Y$  is a vector of  $n$  observations,  $X$  is a matrix consisting of  $p$  predictors,  $\epsilon$  is a vector of i.i.d errors and  $\beta$  is a unknown parameter vector of dimension  $p$ . In the case of  $p > n$  conventional solutions are not feasible and the dimension of  $X$  must be reduced to  $p \leq n$ .

Many attempts have been made to create real time projection models for the quarterly GDP figures based on monthly indicators that have a short publication lag, Zheng & Rossiter (2006). The aim of this chapter is to present variable selection methodology from various fields, and to test its application in generating real-time estimation of the current activity period GDP prior to interest rate decisions. We use a large sample of different monthly indicators which are chosen according to their timing of publication, in order to nowcast quarterly GDP.

Since the number of monthly indicators is larger than the number of observations there has been extensive application of dimension reduction and variable selection techniques via Bridge Equations. These techniques project quarterly data through monthly variables. Many of the previous papers in the field applied Dynamic Factor Analysis (DFA), Angelini et al. (2008) and Banbura et al. (2010), largely follow Giannone et al. (2006). This paper will approach the problem from different directions, comparing five different approaches of dimension reduction and variable selection in order to select the optimal model for projection. We survey a number of methods in this paper to reduce the dimension of data comparing two major schools of practice - dimension reduction conditional and unconditional on the predicted variable.

The unconditional methods focus on Principal Component Analysis and its variants which include DFA, whereas the conditional methods include model selection and prediction through stepwise variable selection, Least Absolute Shrinkage Selection Operator (LASSO), Tibshirani (1996) and the Elastic Net, Zou & Hastie (2005). Throughout the paper the theoretical comparison is complemented by the policy challenge of setting the interest rate of the Israeli economy by the central Bank of Israel.

Out of sample cross validation is evaluated comparing all techniques with a benchmark - the official release of the GDP by the Central Bureau of Statistics of Israel (CBS) - at time of initial release and the current revision for historical data points. We find that the conditional approach outperforms the unconditional approach with relation to predicting the current growth rate of the GDP. Furthermore, we find that the Elastic Net out of sample prediction error is comparable to the official publications' error rate. A distinguishing feature of regularization of linear model is the ability to isolate influential variables which contribute to the real-time assessment. This refinement of the results separate these methods from current ones used in nowcasting and allows the model to be a more comprehensive tool in economic policy decision making. Notable variables that have model inclusion persistence are: The Price of Oil, Employers' Survey, Purchasing Managers' Index, Industrial Production Index, and Employed Persons' Index in Manufacturing of Electronic Motors, Components, and Transport Equipment. Lastly, two case studies are carried out to ascertain the weaknesses of the Elastic Net as a nowcasting method, the case studies include the emergence of the Israel GDP from the economic downturn of 2008-2009 and how the model reacts to unexpected events such as the Second Lebanon War.

Chapter 3 deals with another facet of the continuously accumulation of

data, which has pre-determined decision time points. We look at data that is itself continuously revised, thereby creating a setting where there is no final value of the target process. This creates a new layer of uncertainty to take into account when arriving at a decision: how much will the current data point and thereby the entire series be different within the upcoming periods of time. This can have serious ramifications when a change in policy is undertaken. Such a scenario would be being presented with a certain reality regarding the state of a process and reacting with proper steps, while in retrospect the data representing the same time points in the series was subsequently revised to show a different reality only a number of periods of time later, rendering the policy decision inaccurate, possibly resulting in a negative impact.

We propose a method to estimate the revision process uncertainty. We use this estimation to generate an asymmetrical prediction interval for the upcoming revision of a currently published activity period. These intervals are a function of the maturity of each time period within the current vintage of the growth process. Contemporary economic literature, such as Cunningham et al. (2012), Anderson & Gascon (2009) and Jacobs & Van Norden (2011), model revisions of economic growth through a Kalman Filter while using two major assumptions, one explicit and the other implicit.

The *explicit assumption* is that the published growth can be de-constructed into three parts: the true growth, a time invariant publication bias of a given maturity, and the serially correlated measurement error associated with the publication maturity. The decay rate of the measurement errors is estimated over the full sample and is not a function of the level of maturity. We find that empirically the revisions decay over a long horizon of nearly ten years in the economy we tested, this is twice as long compared to the US and UK

revision decays Cunningham et al. (2012) and Anderson & Gascon (2009). Thus, we do not find it pertinent to model the full path of revisions due to its negligible use for real time policy decision making.

The *implicit assumption* is made in that the revision process has a homogeneous distribution and is not a mixture of a number of different distributions. We postulate that the revision process is a function of the state of the growth process, thereby creating the necessity to model the growth state and estimate model parameters relating to each one. To model this we introduce an additional uncertainty within the high frequency level by defining it as a function of the latent state of the lower frequency. Our hypothesis follows in-line with a similar hypothesis made by Tkacz et al. (2010). They also believed that there is an underlying signal that governs the revisions: ‘In future work, analysts can explore other explanatory variables, as well as understanding whether revisions are likely to be more pronounced in some periods than in others. For example, revisions may be larger around the turning points of business cycles, so in such periods of uncertainty analysts may wish to anticipate large revisions and therefore build larger confidence intervals around their estimates of current GDP growth.’.

The methodology is tested on historical data of the Gross Domestic Product (GDP) of the Israel. We find that there is a significant difference in the size and sign of revisions dependent on the state of the growth rate. More specifically, when the initial publication is in a low growth period, the growth is overestimated and subsequent revisions lower the growth rate, and conversely when the initial publication is in a high growth period the growth is underestimated and the subsequent revisions increase the growth rate.

Combining the added value found in the first two chapters we can formalize a method to derive higher levels of informed policy decision making

in a real time setting. Since the nowcast GDP consists of one out of sample estimate and is by construction an estimated fit of the actual GDP we can safely assume that the same properties of the actual GDP is found in the nowcast series. Continuing this line the prediction interval methodology is applied to the nowcast estimate. This application enhances our estimate and improves the horizon of its effectiveness, where instead of gaining 4 weeks on the official publication we are able to gain estimate 16 weeks using the prediction intervals.

## 2 Policy Driving Methodology

Chapter 4 focuses on how a time sensitive policy problem necessitated the development of a new regularization penalty for linear mixed model in a high dimensional setting. Efficiently and accurately aggregating crowd sentiment to answer probabilistic questions that shape policy is a challenge that was introduced by the lowering of the cost of having a direct and continuous channel to large groups of people, driven by the internet and online social networks. The regularization penalty derived in this chapter is a result of statistical challenges that presented themselves in the Good Judgment Project (GJP), within the Aggregative Contingent Estimation (ACE) Program<sup>1</sup>. The aim of this program is *“to dramatically enhance the accuracy, precision, and timeliness of forecasts for a broad range of event types, through the development of advanced techniques that elicit, weight, and combine the judgments of many intelligence analysts.”*

The randomized control trial design of the GJP, which consists of a hierarchical structure of groups of subjects lends itself to a mixed model to

---

<sup>1</sup>Sponsored by the U.S. Intelligence Advanced Research Projects Activity (IARPA).



take into account the dependency within cohorts and the population fixed effects. The linear mixed model is applied to derive weights for each subject in order to aggregate into one probabilistic prediction to submit for each question. Since the unit of interest is the question and the subject information is training information, a dimensionality problem presents itself. A regularized linear mixed model (LMM) is a solution to such a problem.

The regularization of LMMs and generalized LMMs (GLMM) is a relatively new field of research that is an extension of regularization of generalized linear models (GLM). Advances in the field of model selection and prediction via regularization, using different penalty terms, has forged the ability of a variety of disciplines to classify and model large-scale data. Widely used methods which apply penalties in classification are the Least Absolute Shrinkage and Selection Operator (LASSO), the Adaptive LASSO and the Elastic Net. These methods have predominately been used to classify problems of GLM, discussed in depth in Friedman et al. (2010) and Van de Geer (2008), in which the dependency of the covariance structure is assumed to be independent. This assumption is, in practice, not commonly met and the ability to model such dependencies is integral in fitting the data correctly, such data is modelled using LMMs and GLMMs.

Recent research by Bondell et al. (2010), applied a modified Adaptive LASSO (M-ALASSO), smoothly clipped absolute deviation (SCAD) to LMMs and have produced results of identifying both the random and fixed effects found in data, proving both consistency and an oracle optimality. Model selection within the generalized linear mixed models framework has been discussed in Schelldorfer et al. (2011), Fan & Li (2012), Groll & Tutz (2014), Hui et al. (2016) and Ibrahim et al. (2011). Schelldorfer et al. (2011) and Groll & Tutz (2014) have a drawback that only fixed effects are selected,

while Ibrahim et al. (2011) apply either the SCAD or the ALASSO to each effect. Hui et al. (2016) allow for greater flexibility for different penalty types on the fixed and random effects. It is noteworthy that Ibrahim et al. (2011) tune each penalty term to a different value through the introduction of the IC(q) criterion, a characteristic not found in the other methods.

We propose a new penalty called the linear mixed model Elastic Net, LMMEN, which is better suited for regularization in highly correlated data. The LMMEN allows for regularization of both the sparsity ( $\ell_1$  norm) and grouping ( $\ell_2$  norm) for the fixed and random effects separately. We believe that this method better captures the design of real world data when modelling with linear mixed models, LMMs. Through simulations and the motivating case study we find that the LMMEN out performs comparative methods in three major areas: highly correlated fixed effects data structures, high dimensionality in the fixed effects, i.e.  $p \gg n$  and selection of random effects when the dimension of the covariance matrix is large.

Lastly, chapter 4 contains an extensive appendix detailing the `lmmen` R package Sidi (2017), currently on the CRAN repository, that solves the linear mixed optimization problem with the linear mixed model Elastic Net penalty. We go into greater detail regarding the different types of methods used to solve the optimization problems. We also discuss the implementation of cross validations used in the simulations for both the LMMEN penalty and the other comparative methods. The cross validation of the other methods are functionalities not currently supported in other R packages.

## References

- Anderson, R. G. & Gascon, C. S. (2009). Estimating us output growth with vintage data in a state-space framework. *Federal Reserve Bank of St. Louis Review*, 91(4), 349–69.
- Angelini, E., Camba-Méndez, G., Giannone, D., Reichlin, L., & Rünstler, G. (2008). Short-term forecasts of euro area gdp growth.
- Banbura, M., Giannone, D., & Reichlin, L. (2010). Nowcasting.
- Bondell, H. D., Krishna, A., & Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4), 1069–1077.
- Cunningham, A., Eklund, J., Jeffery, C., Kapetanios, G., & Labhard, V. (2012). A state space approach to extracting the signal from uncertain data. *Journal of Business & Economic Statistics*, 30(2), 173–180.
- Fan, Y. & Li, R. (2012). Variable selection in linear mixed effects models. *Annals of statistics*, 40(4), 2043.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1), 1.
- Giannone, D., Reichlin, L., & Small, D. H. (2006). Nowcasting gdp and inflation: the real-time informational content of macroeconomic data releases.
- Groll, A. & Tutz, G. (2014). Variable selection for generalized linear mixed models by l1-penalized estimation. *Statistics and Computing*, 1–18.

- Hui, F. K., Müller, S., & Welsh, A. (2016). Joint selection in mixed models using regularized pql. *Journal of the American Statistical Association*, (just-accepted).
- Ibrahim, J. G., Zhu, H., Garcia, R. I., & Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, *67*(2), 495–503.
- Jacobs, J. P. & Van Norden, S. (2011). Modeling data revisions: Measurement error and dynamics of true values. *Journal of Econometrics*, *161*(2), 101–109.
- Schelldorfer, J., Bühlmann, P., & Van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics*, *38*(2), 197–214.
- Sidi, J. (2017). *lmmen: Linear Mixed Model Elastic Net*. R package version 1.0.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.
- Tkacz, G. et al. (2010). An uncertain past: Data revisions and monetary policy in Canada. *Bank of Canada Review*, *2010*(Spring), 41–51.
- University of Chicago, U. o. C. (2017). Data science for social good.
- Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 614–645.
- Zheng, I. Y. & Rossiter, J. (2006). *Using monthly indicators to predict quarterly GDP*. Bank of Canada.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

## 2. NOWCASTING ISRAEL GDP USING HIGH FREQUENCY MACROECONOMIC DISAGGREGATES

**Status:** This chapter was submitted to a peer-review journal and is currently under review.

# Nowcasting Israel GDP

Using High Frequency Macroeconomic Disaggregates

Jonathan Sidi<sup>1</sup> and Gil Dafnai<sup>2</sup>

<sup>1</sup>Department of Statistics, Hebrew University of Jerusalem

<sup>2</sup>Department of Economics, Hebrew University of Jerusalem

### **Abstract**

This paper presents a dynamic nowcasting model for estimating the Quarterly GDP in Israel. Currently, monetary policy in Israel is evaluated and updated on a monthly basis. The recent GDP figure is, however, unavailable for monetary policy makers, at the Bank of Israel, at the month following the end of the quarter, due to a six-week lag of the GDP data publication.

The aim of this nowcasting project is to derive "flash" estimates of GDP at a three-week lag, in order to gain three weeks in terms of data availability when updating the interest rate. This is done by utilizing the information contained within a large group of monthly indicators that are available at the relevant date.

Indicator selection, from a pool of these high frequency series, is applied through a variety of dimension reduction techniques. The ability to apply these techniques while conditioning them on the predicted indicator will be examined and discussed in this article.

The Elastic Net is found to be the most comprehensive model selection technique, generating the lowest mean absolute forecast error of only 1.62%. In addition, the Elastic Net successfully captures the timing and magnitude of the 2008-2009 Israeli economic cycle. Notable variables that have model inclusion persistence are: The Price of Oil, Employers' Survey, Purchasing Managers' Index, Industrial Production Index, and Employed Persons' Index in Manufacturing of Electronic Motors, Components, and Transport Equipment.



# 1 Introduction

## 1.1 Background

Central Bank assessments of the current state of the economy play a vital role in the conduct of monetary policy. However, providing an accurate assessment of economic growth in real-time is a challenge that many central banks have to overcome. The challenge is found in the delay between the end of the quarter and the publication of the GDP data. In Israel, quarterly GDP is published at a six-week lag off the end of the relevant quarter.

Many attempts have been made to create real time projection models for the quarterly GDP figures based on monthly indicators that have a short publication lag, e.g. Zheng & Rossiter (2006). The aim of this paper is to present variable selection methodology from various fields, as to test its application in generating real-time estimation of the current quarter GDP. We use a large sample of different monthly indicators which are chosen according to their timing of publication, in order to nowcast quarterly GDP. An indicator will enter the initial set of explanatory variables only if it has at least a value for the first two months of the projected quarter.

Since the number of monthly indicators is larger than the number of observations there has been extensive application of dimension reduction and variable selection techniques via Bridge Equations Baffigi et al. (2004). These techniques project quarterly data through monthly variables. Many of the previous papers in the field applied Dynamic Factor Analysis, e.g. Angelini et al. (2008) and Banbura et al. (2010), largely following Giannone et al. (2006). Table 1 shows the different methods applied in the leading central banks and research centers. This paper will approach the problem from different directions, comparing five different approaches of dimension reduction

and variable selection in order to select the optimal model for projection. Following Klein & Sojo (1990) we use Principal Component Analysis (PCA)

Table 1: Variable Selection Methods Applied in Leading Central Banks

<b>Organization</b>	<b>Method</b>	<b>Number of Variables</b>
Bank of Canada	Simple bridge equation with backward selection	30
ECB (ECARES)	Dynamic Factor Analysis	200
Central Bank of Ireland		41
Central Bank of Portugal		45
Bank de France	Bridge via PCA	Business surveys
This Research	Model Selection via: PCA, SPCA, LASSO, Elastic Net	158

to reduce the dimensionality of the dataset. Since PCA does not set to zero any of the coefficients in the principal components we also apply Sparse PCA, Zou et al. (2006). This method was constructed to improve the inference ability compared to the PCA method. These methods have a serious drawback since they do not incorporate the target variable, i.e. the quarterly GDP, into the dimension reduction procedure. To accommodate this drawback we also examine the application of two variable selection techniques: Least Absolute Shrinkage Selection Operator (LASSO), Tibshirani (1996) and the Elastic Net, Zou & Hastie (2005). Both techniques were developed to address variable selection problems in the field of Bio-Informatics, and both select variables correlated to the response variable by setting constraints on the coefficient of the Least Squares problem.

We find that Elastic Net and LASSO indeed improve the proficiency of nowcasting when compared with both unconditional methods (PCA and SPCA) and univariate multiple regression. In addition the time dependent

Dynamic Factor Model was compared to the other methods and its results were found comparable to the SPCA results. We also find that the price of oil, Purchasing Managers' Index, Employers' Survey, Industrial Production Index, and Employed Persons' Index in manufacturing of electronic motors, components, and transport equipment are the variables with the highest probability to enter the final set of the projection model.

## 1.2 The Problem

In this paper we attempt to apply non-parametric and semi-parametric statistical methodology in order to identify underlying structures in large data sets. The advantage in this approach is two-fold:

1. Minimizing the use of confining structural assumptions on the data.
2. The common properties are deduced from within the data.

Defining the general set  $\Omega^m$  which consists of the relevant monthly time series  $\omega_p^m$  up to month  $m$

$$\Omega^m := \{\omega_p^m | \omega_1^m, \dots, \omega_P^m\}. \quad (1)$$

Due to different publication lags in the data  $\Omega^m$  is redefined as  $\Omega^m \in (\Omega_1^m, \Omega_2^m)$ . Where  $\Omega_1^m$  contains monthly series that have values up to the last month in quarter  $q$ , and  $\Omega_2^m$  contains series in which one month is missing. The missing month in  $\Omega_2^m$  series is forecasted by Holt and Winters Exponential Smoothing. Once the data has no jagged edges, i.e. smoothed, the data is ready to project the GDP of quarter  $q$ . Denoting the GDP projection as:

$$\hat{z}_m^q = Proj[GDP^q | \Omega^m] \quad (2)$$

The selection of the best method for the GDP estimation will be guided by the following criteria:

- The projection  $\widehat{z}_m^q$  should follow the desired property as more data is published:

$$E[(\widehat{z}_{m+1}^q - GDP^q)^2] \leq E[(\widehat{z}_m^q - GDP^q)^2] \quad (3)$$

- The methods will be ranked according to Mean Absolute Forecast Error, MAFE.

## 2 Data Selection Preprocessing

### 2.1 Indicator Selection Methodology

The general set of indicators,  $\Omega^m$ , is compiled of indicators in the general monthly appendix that is used in monthly meetings in the bank in addition to past research Giannone et al. (2006) Angelini et al. (2008) Zheng & Rossiter (2006). In total there are 143 domestic indicators and 15 global indicators<sup>1</sup>. To describe the variables by subject groups Table 2 lists the groups by method/place of collection.

These indicators are characterized by their availability and stability. The availability of monthly data received from the Central Bureau of Statistics (CBS) in comparison to the target meetings, determines the number of indicators included the initial set<sup>2</sup>. Indicator selection occurs twice in the algorithm:

- Structural Criteria

---

<sup>1</sup>The full list of indicators and their description can be found in Table 2 of Web Appendix A

<sup>2</sup>A stationary time line illustrating the indicator publication chronology from the Central Bureau of Statistics can be found in the Appendix Figure 8.

Table 2: Monthly indicators applied to nowcasting by subject group

Subject Group	Number of Indicators
Employed Persons' Index	27
Industrial Production Index	22
Man-Hours Worked Index	21
Non-Domestic Indices	15
Purchasing Manager's Index	11
Employer Survey	10
Retail Trade	8
Revenue index	8
Imports and Exports	8
Housing Indices	7
Bank of Israel	6
Financial and Stock Indices	6
Hotel Occupancy	5
Taxes	3
Consumer Confidence Index	1

The initial selection occurs prior to data transformation fulfilling two conditions:

SC(a). Minimum history of series is 1998m1.

SC(b). At most one month missing from the current quarter.

Defining the resulting data set after the first selection procedure as:

$$\hat{\Omega}^m := \{\hat{\omega}^m \in \Omega^m | \hat{\omega}^m \text{ fullfils SC(a) and SC(b)}\} \quad (4)$$

- General Information Criterion:

The indicator accounts for a high proportion of the total variability in  $\widehat{\Omega}^m$ .

Defining the resulting data set after the second selection procedure as

$$\Theta^m := \{\theta^m \in \widehat{\Omega}^m | \theta^m \text{ fulfills General Information Criteria}\} \quad (5)$$

## 2.2 Data Preprocessing

Data transformation  $\tilde{\omega}^m = f(\widehat{\omega}^m)$  is applied in order to begin the secondary selection procedure given that all the indicators are seasonally adjusted, have the same end point, log-differenced, standardized, and indexed with the beginning of the sample set at 100.

1. Seasonal adjustment is carried out on all series using X-12-ARIMA Findley et al. (1998). The specification file uses the default SARIMA model selection procedure, automatically finds outliers from 1999 to the end of the sample, and the Jewish calendar and trading days are exogenous variables in the model. This is done in order to transform the data to be as similar as possible to seasonally adjusted data of the CBS.
2. Different publication lags for each indicator causes jagged edges in the data, i.e. different end dates from series to series. The Holt and Winters exponential smoother is applied to each indicator that is missing the final month in the current quarter. The additive coefficients of each series are estimated<sup>3</sup>.
3. The log-difference of each seasonally adjusted series is calculated.

---

<sup>3</sup>Estimation was done using the BFGS optimization routine.

4. The percent changes are standardized.

### 3 Methodology

The importance of dimension reduction techniques in modern statistics can be dated back to R.A. Fisher. Fisher is responsible for laying the foundations for modern theoretical and applied statistics. In an article published by Fisher (1922) he defined one of the main goals in statistics as:

”... the objective of statistical methods is the reduction of data. A quantity of data...is to be replaced by relatively few quantities which shall adequately represent...the relevant information contained in the original data.”

Further paraphrasing a Fisher (1925) text, he stated that the variables that are employed as predictors must be chosen without reference to the variable of interest, e.g. current quarter GDP. The article concentrated on the subject of transforming an " $n < p$ " into an " $n > p^*$ " without the dependency of the transformation on the response variable.

In conjunction with Fisher's articles other researchers formulated methods of dimension reduction, including: Adcock (1878), Pearson (1901), Spearman (1904), Hotelling (1933); which are today known as Principal Component Analysis (PCA).

The study of principal components in regression is a case in which the vector of predictors is reduced prior to the regression on the response variable. This is predominantly done in order to mitigate the effects of collinearity and to facilitate model specification by allowing visualization of the regressors in low dimensions, Cook (2009). Additionally it provides a parsimonious set of predictors on which to base interpretation, Cook (2007).

As Fisher stated in his 1924 article, these methods are solely transformations on the explanatory variables. This is the main drawback of dimension reduction when applied to regression. It may be possible to contain the same information in a subset of  $M$  leading principal components as in the population set, but their relationship to the response variable is not addressed. Moreover, an additional drawback is the absence of a conventional method to decide which principal components should be included in  $M$ , this was addressed by Cox (1968):

”A difficulty seems to be that there is no logical reason why the dependent variable should not be closely tied to the least important principal component.”

To overcome these inherent problems in the application of principal components in regression we use sparse regression methodology, Tibshirani (1996) and Zou et al. (2006), to redefine the PCA model as a least squares model, i.e. the Elastic Net.

In the following subsections we define the methods described above beginning with those that are *not conditioned* on the response variable, [3.1.1] and [3.1.2], and then discuss the methods which condition the data reduction on the response variable, [3.2.1]-[3.2.3], a flowchart, Figure 9 in the Appendix, describes how each method is applied in order to identify the final variable set  $\Theta^m$ .

## 3.1 Unconditional Methods

### 3.1.1 Principal Component Analysis

Principal Component Analysis (PCA) is a standard tool in modern data analysis. Its application can be found in a number of diverse fields from



Computer Graphics to Neuroscience. This is because of the simple non-parametric method it uses to extract relevant information from condense data sets. The subsequent section will provide the basic intuition behind PCA, after which its utilization concerning nowcasting will be explained.

The objective of PCA is to find a linear transformation that reduces the dimension of a multivariate sample  $\mathbf{X}_{(n \times p)}$  defined as:  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p$  into  $\hat{\mathbf{X}}_{(n \times q)}$  where  $q < p$ . The transformed set has the following desirable properties:

1. The elements of  $\hat{\mathbf{X}}$  are uncorrelated.
2. Each element in  $\hat{\mathbf{X}}$  should account for as much of the combined variance of the elements in  $\mathbf{X}$  as possible.  $\hat{\mathbf{X}}$  is selected so as to minimize redundant information in the latent variables by maximizing variance of the relevant variables in the data, thereby minimizing information loss.

Formally, the objective of PCA is to minimize the target function  $F(\cdot)$ , where  $F(\cdot) = c_i^t \mathbf{R} c_i$  subject to constraints:

$$\max \quad c_i^t \mathbf{R} c_i \quad (6a)$$

$$s.t.(1) \quad \|c_i\|_2 = 1 \quad (6b)$$

$$s.t.(2) \quad c_i^T \mathbf{C}_{i-1} = (0, 0, \dots, 0) = 0_{i-1}^t \quad (6c)$$

In this problem we define  $\mathbf{R}$  as the standardized covariance or correlation matrix where the vectors  $c_i$  are the solution to the maximization problem. The formulation in (6) gives an intuitive insight into the main purpose of PCA, which is to find the directions which maximize the variance in  $\mathbf{X}$ . Notice that (6b) is necessary in order to ensure the problem to have a finite

solution and (6c) assures that each successive solution  $c_i$  is orthogonal to all the previous solutions,  $\mathbf{C}_{i-1}$ .

A practical solution to (6) is through the use of Singular Value Decomposition (SVD). SVD allows us to take any matrix  $\mathbf{X}$  and decompose it into the eigenvalues and eigenvectors. Defining the SVD of  $\mathbf{X}$  as:

$$\mathbf{X}\mathbf{V} = \mathbf{\Lambda}\mathbf{U} \quad (7a)$$

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t \quad (7b)$$

We define two orthogonal matrices  $\mathbf{U}$  and  $\mathbf{V}$ , and a diagonal matrix  $\mathbf{\Lambda}$ . The construction of  $\text{diag}(\mathbf{\Lambda})$  is of the form  $(\sigma_1 \dots \sigma_p, 0 \dots 0)$ , where  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p$ . The columns of  $\mathbf{U}$  are called principal components of unit length, and the columns of  $\mathbf{V}$  are the corresponding loadings of the principal components, i.e. each  $\mathbf{V}_i$  are eigenvectors of matrix  $\mathbf{R}$ , defined  $\mathbf{R} = \mathbf{X}^t \mathbf{X}$ . By applying some linear algebra:

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t \quad (8a)$$

$$\mathbf{X}^t \mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{U}^t \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t = \mathbf{V}\mathbf{\Lambda}^2 \mathbf{V}^t \quad (8b)$$

$$\text{tr}(\mathbf{X}^t \mathbf{X}) = \text{tr}(\mathbf{V}\mathbf{V}^t \mathbf{\Lambda}^2) \quad (8c)$$

$$\text{tr}(\mathbf{X}^t \mathbf{X}) = \text{tr}(\mathbf{R}) = \sum_{i=1}^p \lambda_i^2 \quad (8d)$$

In (8d) we conclude that  $\lambda_i$  are eigenvalues of  $\mathbf{R}$ , furthermore they represent the variance of  $\mathbf{R}$  along a given  $\mathbf{V}_i$ . This is due to the fact that  $\|\mathbf{X}\mathbf{v}_i\|^2 = \lambda_i \equiv \sigma_i^2$ . After we solve for the eigenvectors we can use them to change the base of  $\mathbf{X}$  to the orthogonal base  $\hat{\mathbf{X}}$ :

$$\mathbf{V}^t \mathbf{X}^t = \hat{\mathbf{X}} \quad (9)$$

In regression there are many cases in which there are more variables,  $p$ , than observations,  $n$ . In these cases, PCA is used in order to create new variables, the latent scores  $\hat{\mathbf{X}}$  as described above, which are used as the observed variables in the regression. This is done to decrease the effects of collinearity and simplify the interpretation of the regressors in the model.

We test three methods in the use of principal components in multivariate regression. The first uses the leading principal components as independent variables in regression, while the other two apply variable selection from within the components.

### Classic Approach

Defining the subset  $\Theta^m$  consisting of  $q$  principal components. The number of components may be set equal to the number of components that their corresponding  $\lambda$  is greater than some  $\lambda_0$ . The level of  $\lambda_0$  used in this method was set at 0.7 Jolliffe (1972). The use of the components in nowcasting is less favorable because the components used are linear combinations of all the variables in  $\hat{\Omega}^m$ . Using such a large set for policy decisions is not practical.

### Two Component Norm

In this procedure the norm of the first two principal components from each variable is calculated and then sorted in ascending order. This is a naive variable selection procedure because it only utilizes the first two components. The maximum variation that can be explained is  $\frac{\sum_{i=1}^2 \lambda_i}{\sum_{i=1}^p \lambda_i}$ . Within this portion of the total variance we are choosing the subset of variables that have the largest weights.

---

**Algorithm 1** Two Component Norm
 

---

- |    |                                       |
|----|---------------------------------------|
| 1: | Define $\tilde{V} = [v_1 v_2]$        |
| 2: | Calculate by rows $Q = \ \tilde{V}\ $ |
| 3: | Sort Ascending Q                      |
| 4: | $\Theta^m = \bigcup_{j=1}^8 Q_j$      |
- 

**Iterated Component**

This variable selection method, introduced by Jolliffe (1972), forms a subset of  $K$  independent variables for multivariate regression using principal components. A subset of the first  $K$  components is formed by setting a constraint  $\lambda_k \geq \lambda_0$ , for a given  $\lambda_0$ . The same level of  $\lambda_0$  was used as in the classic approach (0.7). The variable with the largest coefficient in  $K_1$ , the component with the largest eigenvalue, is placed in subset  $\Theta^m$ . Then iteratively one variable is chosen which is associated with the remaining  $K-1$  components under consideration and which has not already been placed in  $\Theta^m$ .

---

**Algorithm 2** Iterated Component
 

---

1:	Define $\lambda_0$
2:	if $\lambda_i \geq \lambda_0$ then $PC_i \in K, i = 1 \dots P$
3:	$v_j = \arg \max (K_1)$
4:	$v_j \in \Theta^m$
5:	for $k=2$ to $\text{length}(K)$
6:	$v_j = \arg \max (K_k)$
7:	if $v_j \notin \Theta^m$
8:	then add to $\Theta^m$
9:	else check next largest $v_j$
10:	end if
11:	next

---

### 3.1.2 Sparse PCA

In the previous section we discussed different methods to identify variables after applying PCA to the data set. The main drawback of PCA is that the loadings are inherently nonzero. This makes it difficult to interpret PCs when applying it to large data sets. In the following section we will describe a technique that produces modified principal components, i.e. sparse loadings. This will be done by formulating the PCA optimization as a regression-type optimization problem, and imposing two exterior penalty functions on the regression coefficients.

The application of this algorithm on principal components was first introduced by Zou et al. (2006). The exterior penalty functions are defined as:

1.  $L_1$ -norm constraint:  $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$

This constraint effectively act as a scaling parameter on the solution

to the minimization problem.

$$2. L_2\text{-norm constraint: } \|\beta\|^2 = \sum_{j=1}^p (\beta_j)^2$$

The Ridge penalty/Tikhonov regularization is a well known method for reducing variability of coefficients in a regression through the bias-variance trade off. This constraint allows for highly correlated coefficients to be grouped together.

The PCA problem (7) can be viewed as a simple regression problem with a ridge penalty, where  $X_i$  is the  $i$ -th row vector,  $\alpha = \alpha_{(p \times k)}$ ,  $\beta = \beta_{(p \times k)}$ , and  $\forall \lambda > 0$  we derive:

$$(\hat{\alpha}, \hat{\beta}) = \arg \max_{\alpha, \beta} \sum_{i=1}^n \|\mathbf{X}_i - \alpha \beta^t \mathbf{X}_i\|^2 + \lambda \sum_{j=1}^k \|\beta_j\|^2 \quad (10a)$$

$$s.t. \quad \alpha \alpha^t = I_k \quad (10b)$$

Then we get  $\hat{\beta}_j \propto \mathbf{V}_i$  for  $j = 1 \dots k$ , where  $I_k$  is a square identity matrix of dimension  $k$ . The intuition behind (10) is that if we set  $\alpha = \beta$  then  $\sum_{i=1}^n \|\mathbf{X}_i - \alpha \alpha^t \mathbf{X}_i\|^2$  then we get an alternative formulation of the standard PCA problem that gives the same solution, under orthonormal constraints. As in the case of the nowcasting problem setting, " $p > n$ ", requiring  $\lambda > 0$  ensures (10) yields the exact PCA solution.

Finally defining the SPCA problem by adding the  $L_1$ -norm constraint to (10) in order to obtain sparse loadings.

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \arg \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \underbrace{\sum_{i=1}^n \|\mathbf{X}_i - \boldsymbol{\alpha} \boldsymbol{\beta}^t \mathbf{X}_i\|^2}_{\text{PCA}} + \lambda \sum_{j=1}^k \|\beta_j\|^2 + \underbrace{\sum_{j=1}^k \lambda_{1,j} \|\beta_j\|_1}_{\text{Coefficient Calibration}} \quad (11a)$$

$$s.t. \quad \boldsymbol{\alpha} \boldsymbol{\alpha}^t = I_k \quad (11b)$$

The different  $\lambda_{1,j}$  are allowed for penalizing the loading of different PCs, and the ridge penalty is constant over all  $k$  components. While Sparse PCA is a marked improvement in the ability to interpret the loadings in comparison to PCA the method lacks the qualities we are aiming for as a tool for model selection in nowcasting. The main drawbacks to this method are:

- To form sparse components where no other variable selection method is needed causes the proportion of  $\frac{\text{Variance Explain}_{\text{SPCA}}}{\text{Variance Explain}_{\text{PCA}}}$  to decrease beyond accepted levels.
- The solution to the optimization problem is not conditioned on the variable of interest.

## 3.2 Conditional Methods

### 3.2.1 Univariate Regression

The production of real-time forecasts from univariate equations by regressing current-quarter GDP on all variables in a general set has been used extensively in central banks e.g the Central Bank of Canada.

Although this method is quite simplistic, it produces low forecasting errors in the absence of full quarter data, we have included this method as a benchmark of the methods that condition the subset selection on the GDP.

For each variable in the data set a univariate regressions is run on the current quarter GDP, denoted as  $GDP^m$ .

$$GDP^m = c + \beta_i \hat{\omega}_i^m, \quad \forall \hat{\omega}_i \in \hat{\Omega}. \quad (12)$$

A Subset of the 25 variables with the lowest AIC is taken, after which a stepwise backward regression is applied with the subset as the independent variables and the GDP as the dependent variable. The stopping criterion, removal p-value, in the selection is 10 percent.

---

**Algorithm 3** Univariate Regressions

---

- |    |  |
|----|--|
| 1: | Run $GDP^m = c + \beta_i \hat{\omega}_i^m + \epsilon, \quad \forall \hat{\omega}_i \in \hat{\Omega}$   |
| 2: | $Q = AIC_i$ sorted in descending order   |
| 3: | $\hat{Q} = \bigcup_{j=1}^{25} Q_j$   |
| 4: | Run Stepwise Backward Regression on<br>$GDP^m = c + \sum_{i=1}^{25} \beta_i \hat{Q}_i + \epsilon$ <p>The variables that are in the final regression<br/>make up the subset <math>\Theta</math></p> |
- 

While improving the simple regression by selecting a subset of the original data the stepwise method activated on a subset of variable has two drawbacks:

1. The accepted stepwise methods used do not calculate all the models possible if there are more than eight possible variables in the model. Jolliffe (1972)
2. Prediction stability is a problem because small changes in the data can result in very different models being selected.



### 3.2.2 Regression with a Tuning Parameter

Given the ordinary least squares (OLS) formulation where target variable  $\mathbf{Y}$ <sup>4</sup> and the estimation compromised of the training data  $\mathbf{X}$ , the target of OLS is to minimize the residual square error loss function.

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^N (y_i - x_i^t \beta)^2 \quad (13)$$

There are two central reasons why an analyst would be unsatisfied with the OLS estimators.

#### Prediction accuracy

OLS estimates are defined to have zero bias (BLUE) but this causes the variance of the model to increase, therefore prediction accuracy can be improved by setting some of the coefficients to zero.

#### Interpretation

In the case of a large number of predictors, we often would like to determine a canonical subset that exhibits the strongest effects, increasing the ability to draw inference from the results.

In an attempt to address the issues with variable selection procedures the LASSO (Least Absolute Selection and Shrinkage Operator) was introduced by Tibshirani (1996). The LASSO solves the minimization problem (13):

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^N (y_i - x_i^t \beta)^2 \quad (14a)$$

$$st \sum_j |\beta_j| \leq t \quad (14b)$$

---

<sup>4</sup>The target variable is centered prior to optimization.

The same penalty function was used in the case of Sparse PCA,  $L_1$ -norm [1]. The LASSO tends to shrink the OLS coefficients toward zero, and setting some exactly to zero leaving only the most important ones. This often improves prediction accuracy, while trading off decreased variance for increased coefficient bias.

### 3.2.3 Regression with a Tuning Parameter and a Grouping Penalty

As we have seen in the previous section penalizing a regression with the  $L_1$ -norm improves OLS in sense of variability of coefficients and sparsity. Although it has shown success in many situations, it has limitations:

1. Where " $p > n$ " the LASSO selects at most  $n$  variables.
2. Where " $n > p$ ", if there are high correlations between predictors, it has been empirically observed that the prediction performance of the LASSO is dominated by ridge regression Tibshirani (1996).<sup>5</sup>
3. If there is a group of variables among which the pairwise correlations are very high, then the LASSO selects only one variable from the group and does not differentiate which one is selected.

Concerning nowcasting the third limitation theoretically makes the LASSO an inferior variable selection method. This problem was first addressed by data analysts who worked with micro arrays, where the number of variables is extremely high and grouping is a desirable property. Zou et al. (2006) proposed the Elastic Net, which integrates the  $L_1$ -norm and  $L_2$ -norm penalties together thus gaining the desirable property of grouping. The Elastic Net

---

<sup>5</sup>It should be noted that this shortcoming is irrelevant to this paper.

is a convex combination of the ridge penalty and the LASSO penalty. The Elastic Net solves the following problem Friedman et al. (2010):

$$\min_{\beta \in \mathbb{R}^{p+1}} \left[ \frac{1}{2N} \sum_{i=1}^N (y_i - x_i^t \beta)^2 + \lambda P_\alpha(\beta) \right] \quad (15a)$$

where

$$P_\alpha(\beta) = (1 - \alpha) \frac{1}{2} \|\beta\|_{l_2}^2 + \alpha \|\beta\|_{l_1} = \sum_{j=1}^p \left[ \frac{1}{2} (1 - \alpha) \beta_j^2 + \alpha |\beta_j| \right] \quad (15b)$$

## 4 Results

### 4.1 Israel GDP: Descriptive Analysis

Prior to presenting the results of each method described above we will briefly discuss the descriptive attributes of the quarterly seasonally adjusted GDP released by the CBS. Additionally, we will mark points of interest that will be expanded upon in a case study following the results.

Table 3: Descriptive Statistics: Israel GDP

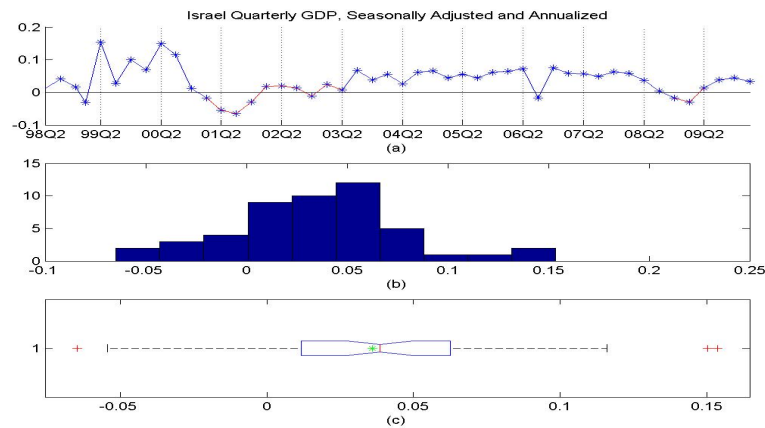
Israel GDP 1998q1-2010q1			
Seasonly Adj. and Annual Percent Change			
Mean	Median	Std	KS (Pvalue)
3.60%	3.85%	4.50%	47.42%

During the sample period (1998q1-2010q1), the economic climate in Israel which was characterized by steady growth accompanied by two business cycles that contributed to the variation, as seen in Figure 1(a). Furthermore, we see that the published GDP is not distributed normally, as we reject the

null hypothesis tested in the Kolmogorov-Smirnov (KS) test <sup>6</sup>. In addition, there is evidence of positive skewness and elevated levels of kurtosis to the distribution, as seen in Figure 1(b). Using the boxplot in Figure 1(c) we locate possible outliers of the series at the peaks (1999q2 and 2000q2) and the gully (2001q3) of the high tech bubble.

A case study will concentrate on the recessions marked in red<sup>7</sup> in Figure 1(a), to test how well the model reacts to external shocks to the economy. In addition, as part of a more general test of robustness of the algorithm, we will focus on how the model reacts to the data with and without the Second Lebanon War (2006q3). This will test if the event can be treated as an outlier of the published series.

Figure 1: Analysis of Distribution Properties of the Published Quarterly GDP.



<sup>6</sup>The null distribution of this statistic is calculated under the null hypothesis that the sample is drawn from the reference distribution, in this case the Gaussian distribution.

<sup>7</sup>The NBER method was used for recession identification.

## 4.2 Constraint Levels

The selection of the constraint levels in the optimization problems applied in this paper is paramount in producing results with low error rates and correct levels of sparsity. In this section we will discuss the dynamic constraint level selection procedures applied in the general algorithm.

As part of the algorithm which solves optimization problem<sup>8</sup>, we select values of  $\alpha$  in the range [0.1, 0.2, ..., 1.0]. For each value of  $\alpha$  the algorithm generates solutions to the optimization problem between two extremes on the range of  $\lambda$ ,  $[\lambda_L(\alpha), \lambda_H(\alpha)]$ , where:

$$\|\beta(\lambda)\| = \begin{cases} \|\beta_{OLS}\| & \lambda < \lambda_L(\alpha) \\ 0 & \lambda > \lambda_H(\alpha) \end{cases}$$

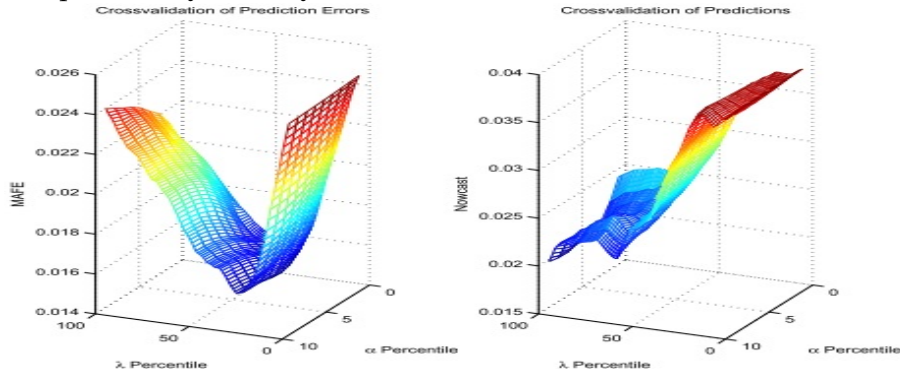
It was found that the predictions of the GDP in the range of  $[\lambda_L(\alpha), \lambda_H(\alpha)]$  are locally robust with relation to close values of  $\lambda$ . On the other hand, the variability increases when comparing the predictions between the quantiles of  $[\lambda_L(\alpha), \lambda_H(\alpha)]$ . Figure 2 shows the crossvalidation results of predictions and prediction errors (MAFE), for each level of  $\alpha$  and  $\lambda(\alpha)$ . This procedure is applied to determine which constraint levels give the best out of sample results each time the general algorithm is run.

Additionally, as discussed in the previous section, the SPCA problem can also be reformulated as an Elastic Net problem. The values of  $\alpha$  and  $\lambda$  are set automatically by the algorithm<sup>9</sup> in order to solve for a maximum 20 non-zero coefficients per component. This amount of non-zero coefficients is sufficient to run the subsequent variable selection methods, i.e. Classic Approach, Two

<sup>8</sup>The glmnet package which solves the Elastic Net optimization problem is available for both Matlab and R can be found at <http://www-stat.stanford.edu/~tibs/lasso.html>

<sup>9</sup>The LARS Matlab package, found in the same link referred to the glmnet, was used to calculate the solution for the SPCA problem.

Figure 2: Crossvalidated Levels of Predictions and Forecast Errors, Elastic Net Sample 2004Q2-2010Q1



Component Norm and Iterated Component, and stepwise regressions applied in the unconditional methods.

### 4.3 Main Results

In order to assess the goodness of fit of our models we have conducted a rolling regression of 24 periods, beginning at 2004Q2. We then calculated the out of sample projection for each period accompanied by an identical set of statistics for each of the methods. The statistics included are: Standard error of the projection (S.E), Adjusted  $R^2$ , Akaike Information Criterion (AIC), Root Mean Square Error (RMSE), Durbin Watson statistic (DW) and the Kolmogorov-Smirnov (KS) test. Table 4 summarizes the results of the different methods by averaging each statistic over the 24 periods on which we conducted an out of sample projection. In addition, we constructed a series called  $CBS_{first}$  which consists of the first vintage quarterly GDP data published by the CBS. This will serve as our control series to compare out of sample results. Figure [3] shows a comparison of the projected and published GDP series within the conditional and unconditional methods.

Figure 3: Comparison of Out of Sample Projection

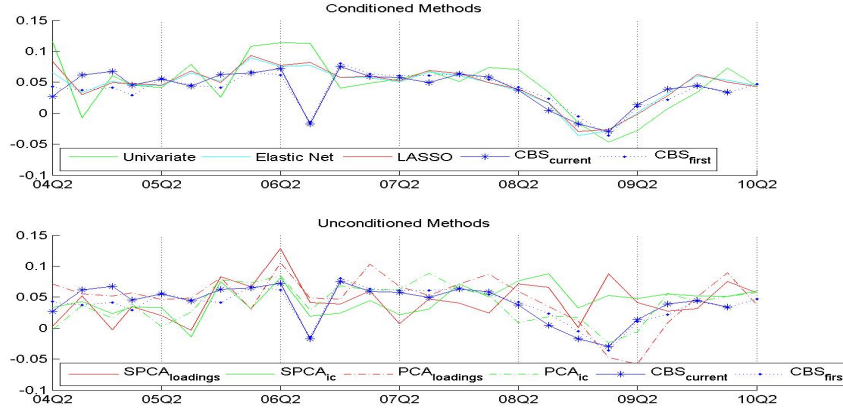


Table 4: Statistics of Goodness of Fit of the Selection Methods

Method		S.E.	$R^2_{Adj}$	AIC	RMSE	DW	KS test
Conditional	Univariate	3.0%	58.9%	-125.6	2.9%	2.5	86.9%
	LASSO	3.5%	51.3%	-123.6	3.3%	2.6	72.6%
	Elastic Net	3.6%	50.8%	-123.2	3.3%	2.7	81.3%
Unconditional	IC	3.6%	42.9%	-112.9	3.4%	2.3	70.0%
	TCN	4.0%	29.5%	-104.6	3.8%	1.8	84.7%
	PCA: Loadings	3.5%	46.8%	-114.8	3.3%	2.5	73.5%
	SPCA: Loadings	3.4%	48.8%	-117.6	3.2%	2.4	74.0%

Table 4 may be quite misleading at first, since it displays data from the rolling regression which by construction takes into account the in-sample projections. Therefore, even though this table is quite informative regarding the goodness of each model it does not provide an answer to the most important question for this paper: Which selection method best projects the quarterly GDP?

In Figure 4 we show the distribution of the absolute projected errors for each method. We clearly see that the Elastic Net, followed by the LASSO,

provides the best real-time projection. Figure 4 is reinforced by Table 5 which shows Wilcoxon Rank Sum test, which tests for equal forecast performance between competing methods. This is done by a non-parametric test of the hypothesis that a pair of series have equal location. From this table we conclude that the Elastic Net and the LASSO are similar to first release of the GDP in the context of performance, while the other methods reject this hypothesis. Furthermore, this table gives a simple method to compare all the methods to each other.

Another question that is of interest to us is to what extent the density of the results from the different models is similar to the actual GDP, and even more importantly, whether the out of sample predictions are unbiased. We find that out of sample density of each projection method of the LASSO and the Elastic Net are slightly biased, though comparatively less than the other methods. In contrast to nowcasting models applied today in most central banks we did not apply time dynamics in this research. We justify this decision by testing for serial correlation in the projected GDP of the Elastic Net and LASSO. We apply to the residuals the Breusch-Godfrey Serial Correlation LM test. We find that the null hypothesis of no serial correlation was not rejected in either final model<sup>10</sup>.

### 4.3.1 Importance of Different Series

Table 1 in Web Appendix A shows the probability of each variable to enter each model. We find that the conditional methods present very high consistency over time, in contrast to the unconditional methods which showed many substitutions of different variables over time. This implies that there are strong correlations between several variables and GDP, however these

---

<sup>10</sup>Results of the LM test can be found in Table 6 the Appendix.



Figure 4: Out of Sample Forecasts Performance Comparison

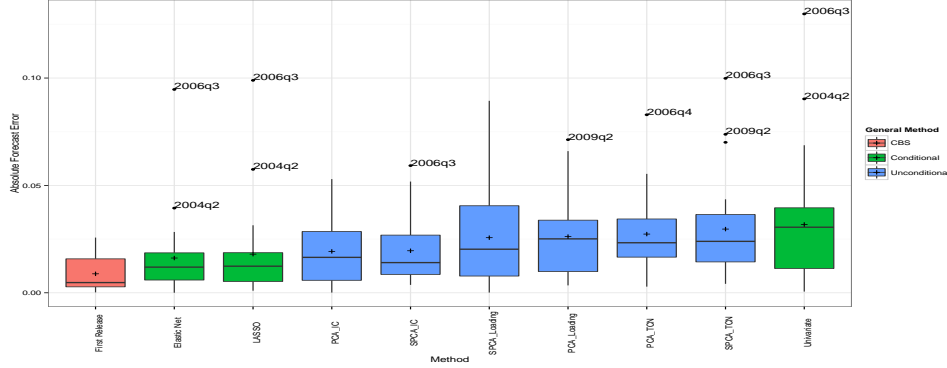


Table 5: Test for Equal Forecast Performance Between Methods

		Wilcoxon Rank Sum Test (Pvalues)								
						PCA		SPCA		
		First Release	Elastic Net	LASSO	Univariate	Loading	IC	TCN	Loading	IC
PCA	Elastic Net	<b>0.13</b>	-	-	-	-	-	-	-	-
	LASSO	<b>0.05</b>	<b>0.74</b>	-	-	-	-	-	-	-
	Univariate	0.00*	0.02*	0.03*	-	-	-	-	-	-
	Loading	0.00*	0.02*	0.04*	0.62	-	-	-	-	-
	IC	0.01*	0.38	0.55	0.15	0.17	-	-	-	-
SPCA	TCN	0.00*	0.00*	0.00*	0.98	0.68	0.07	-	-	-
	Loading	0.00*	0.00*	0.01*	1.00	0.69	0.09	0.99	-	-
	IC	0.01*	0.35	0.56	0.15	0.23	0.98	0.04*	0.07	-
	TCN	0.01*	0.15	0.27	0.51	0.62	0.49	0.41	0.37	0.55

\*Reject hypothesis that the difference of a location parameters of each pair is 0 at 0.95

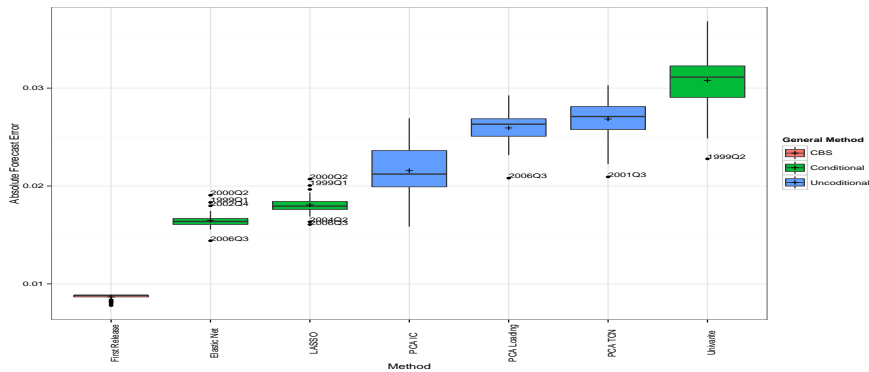
variables do not necessarily account for a large part of the data variability over time. In the conditional methods we find that the variables that enter the final set  $\Theta^m$  of the projection model can be categorized into three groups; a) Domestic Indicators b) Market Expectations Indices c) Global Variables. The variables with the highest probability of entering  $\Theta^m$  include: Price of Oil, Purchasing Managers' Index, Employers' Survey, Industrial Production Index, and Employed Persons' Index in Manufacturing of Electronic Motors, Components, and Transport Equipment. Figures 2(a)-2(c) in Web Appendix

B show the size of the coefficients of all the variables chosen in each period of the rolling regressions.

### 4.3.2 Robustness of the Algorithm

Finally we test if the results from each method is robust in its prediction. This is done by applying the Jackknife technique to the sample. We use the Jackknife to estimate the bias and the variance of the absolute error of each method, by leaving out one observation at a time from the sample set. From Figure 5 we see the conditional methods are more stable than the unconditional ones. Within the conditional methods the over-sensitivity to changes in the data in the univariate is highlighted, the Elastic Net has both lower mean absolute errors and variation compared to the LASSO. Detailed results of the each method within each period chosen in the Jackknife can be found Figures 4(a)-4(d) in Web Appendix B.

Figure 5: Mean Absolute Errors in Jackknife Procedure



## 4.4 Inference

While conventional econometrics is based on structural models, which by construction produce unbiased estimates, the methods utilized in this paper

break those assumptions by adding a penalty to the minimization problem. It has been shown in the previous section that the out of sample predictions produced by the LASSO, and its general form the Elastic Net, out perform simple regression and classic dimension reduction techniques. The question left unanswered is the ability to inference using the coefficients produced in sparse regression. The constraints used in the Elastic Net were formulated for variable selection and do not solve the inherent multicollinearity problem found in OLS. Thus, reaching conclusions as to the effect of each variable on the response variable may be tenuous. Conclusions that can be extracted from the Elastic Net are the characteristics of the variables chosen in the final subset and the proportion which each variable contributes to the forecast level. To facilitate economic policy decisions a comparison can be conducted to understand changes in composition to the final subset and the different magnitude of the persistent variables throughout the evolving economic business cycle.

#### **4.5 Case Study: The Effect of External Shocks on Model Accuracy**

There are two prominent episodes during our sample period, that exemplify abnormal economic activity. We examine these periods in order to assess and analyze algorithm performance in unusual times. The first is the Second Lebanon War, 2006Q3, in which the largest absolute error from the published GDP occurred and the previous global economic crisis 2008Q2-2009Q4. We discuss the subject of short comings of the algorithm and if they have economic explanations or if they are methodological errors.

#### 4.5.1 Anomalies in the Data: The Second Lebanon War

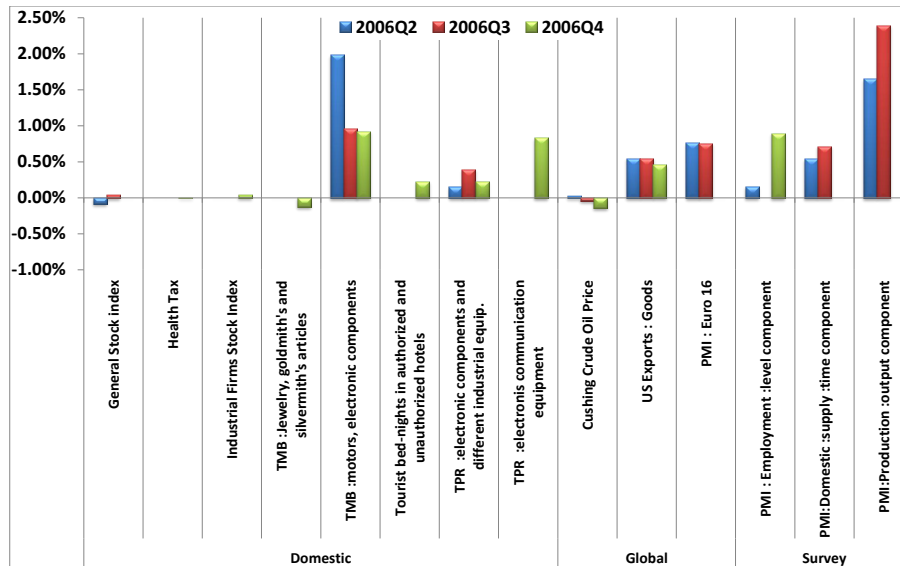
The GDP growth rates which preceded and follow the Second Lebanon War, 2006Q3, were 7.5% and 7.8%, respectively. During this period there was steady global economic growth. Expectations of conflict in the northern border were minimal, and consequently the macroeconomic impact of the Second Lebanon War was unexpected and instantaneous. There was minimal impact to the growth levels of all the major indicators of economic activity: private consumption (3.4%), government consumption (13.6%), fixed capital formation (26.1%), unemployment (6.9%), business sector labor hours (-1.2%). It is evident that the abrupt deviation from steady growth in the GDP (-1.5%) was the consequence of an unexpected drop in inventories.

This conclusion is not surprising when recalling the immediate impact that the war had on the Israeli economy and in particular the labor market. During the 34 days of conflict the northern region was paralyzed due to the constant mortar shelling, and a major portion of the workforce was enlisted to reserve duty. These factors led to a reduction in production capabilities of the Israeli economy, reflected in the steep decline of 12% in the utilization of machinery and equipment index during the quarter. A review of the variable contribution, Figure 6, in the LASSO model reveals that compared to the previous period the level of the variables have a similar behavior which we see in the National Accounts data, i.e. not reflecting the unexpected drop in the GDP. A variable that could have captured the shock is Number of Tourist Bed-Nights, which exhibited an 8.4% drop in 2006Q3. Internal research in the Bank of Israel, Menashe and Sharhabani, has found this variable to be highly correlated with the level of security concerns in Israel. However, since this variable does not show high predictive ability during steady economic activity, it was not selected by the LASSO when it could have actually been

most indicative.

This special case provides an important insight to the importance of variable selection to the general set of indicators. The current general set is almost exclusively comprised of uses related variables. This in turn causes the algorithm to be insensitive to changes in the GDP which are induced by short term and unexpected factors, such as local conflicts or natural disasters.

Figure 6: Variable Contribution for Nowcast 2006Q2-2006Q4 (LASSO)



#### 4.5.2 Unmatched Market Expectation: Emerging out of Crisis

Compared to the global markets Israel has encountered more of an economic slowdown than a financial and real estate crisis in the past two years (2008Q2-2009Q4). Nevertheless, Israel did suffer from four consecutive quarters of comparatively low economic activity, two of which included contraction of the GDP. Analysis of the performance of the model in signaling the entrance and the emergence from an economic slowdown is prudent to understand how the algorithm reacts to economic instability. The algorithm captured the timing, magnitude, and depth of the downturn. While capturing the timing of the beginning of the recovery it miss-timed the end of the recovery. We will briefly discuss the reasons we believe the persistent high growth rate continued through 2009Q4 (5.9%) while the published GDP tailed off (4.4%).

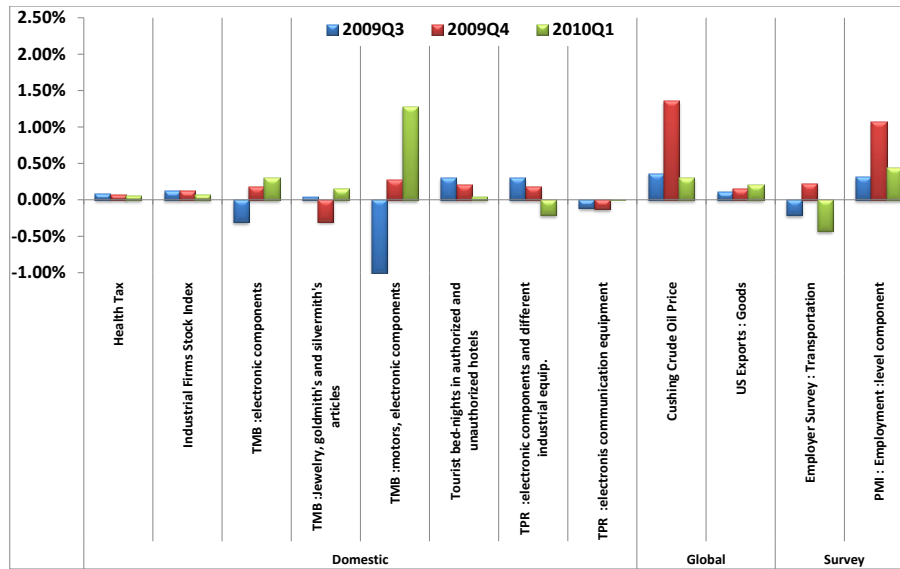
Level shifts in time series are difficult to capture, one may argue that this contributed to the temporary inaccuracy of the model. However, the model performed fairly well during the level shifts in the aforementioned global crisis. In addition, a closer examination of the variables that were chosen, Figure 7, reveals that even though the variables with the largest coefficients actually decreased during this period, the Purchasing Managers' and the Employers' Indices had higher than expected levels of optimism in 2009Q4, causing the extended increase in projected growth<sup>11</sup>.

This case study reveals a source for potential inaccuracies in the projected data, due to the fact that it uses market expectations and not only market data. This drawback shall be examined in future research in order to improve model accuracy.

---

<sup>11</sup>A full decomposition of coefficient levels is available in Figures 1-2, Web Appendix B.

Figure 7: Variable Contribution for Nowcast 2009Q3-2010Q1 (LASSO)



## 5 Conclusions

Policy decision making in central banks is dependent on real-time data analysis as it is published. The ability to produce precise nowcasts through canonical models has evolved with the methodological progress of model selection techniques. Advances in different fields of research have improved model selection for large scale problems. These advances in nowcasting have yet to be fully utilized.

This paper compared model selection techniques applied in leading central banks today with a new method, the Elastic Net. The application of nowcasting with the Elastic Net to the Israel GDP yielded more precise and stable results. Moreover, the dynamic nature of the model allows it to adapt to shocks in the economy producing a more robust model.

A distinguishing feature of the Elastic Net is the ability to isolate influential variables which contribute to the real-time assessment. This refinement of the results separates this method from current ones used in nowcasting and

allows the model to be a more comprehensive tool in economic policy decisions. Finally, this research highlighted the contribution of advanced data mining techniques in a policy driven economic setting. Further development and adaptation of the inference ability of these techniques could broaden the insight into many structural econometric models applied today.

## References

- Adcock, R. J. (1878). A problem in least squares. *The Analyst*, 5, 53–54.
- Angelini, E., Camba-Méndez, G., Giannone, D., Reichlin, L., & Rünstler, G. (2008). Short-term forecasts of euro area gdp growth.
- Baffigi, A., Golinelli, R., & Parigi, G. (2004). Bridge models to forecast the euro area gdp. *International Journal of forecasting*, 20(3), 447–460.
- Banbura, M., Giannone, D., & Reichlin, L. (2010). Nowcasting.
- Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1), 1–26.
- Cook, R. D. (2009). *Regression graphics: Ideas for studying regressions through graphics*, volume 482. Wiley-Interscience.
- Cox, D. (1968). Notes on some aspects of regression analysis. *Journal of the Royal Statistical Society. Series A (General)*, 265–279.
- Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C., & Chen, B.-C. (1998). New capabilities and methods of the x-12-arima seasonal-adjustment program. *Journal of Business & Economic Statistics*, 16(2), 127–152.



- Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222, 309–368.
- Fisher, R. A. (1925). The influence of rainfall on the yield of wheat at rothamsted. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213, 89–142.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1), 1.
- Giannone, D., Reichlin, L., & Small, D. (2006). Nowcasting gdp and inflation: the real-time informational content of macroeconomic data releases.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *The Journal of educational psychology*, 498–520.
- Jolliffe, I. (1972). Discarding variables in a principal component analysis. i: Artificial data. *Applied statistics*, 160–173.
- Klein, L. & Sojo, E. (1990). Combinations of high and low frequency data in macroeconometric models. In *Economics in Theory and Practice: An Eclectic Approach* (pp. 3–16). Springer.
- Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559–572.
- Spearman, C. (1904). " general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2), 201–292.

- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.
- Zheng, I. Y. & Rossiter, J. (2006). *Using monthly indicators to predict quarterly GDP*. Bank of Canada.
- Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.
- Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2), 265–286.

## A Appendix

Table 6: Test for Serial Correlation

Breusch-Godfrey Serial Correlation LM Test	Coefficient	SE	t Stat	Prob.
EN	0.017	0.092	0.18	0.86
RESID(-1)	-0.230	0.225	-1.02	0.32
RESID(-2)	0.027	0.223	0.12	0.90
LASSO	0.025	0.098	0.26	0.80
RESID(-1)	-0.245	0.227	-1.08	0.29
RESID(-2)	-0.002	0.223	-0.01	0.99

Figure 8: Central Bureau of Statistics Israel Series Publication Time Line  
Anchor 2010Q1

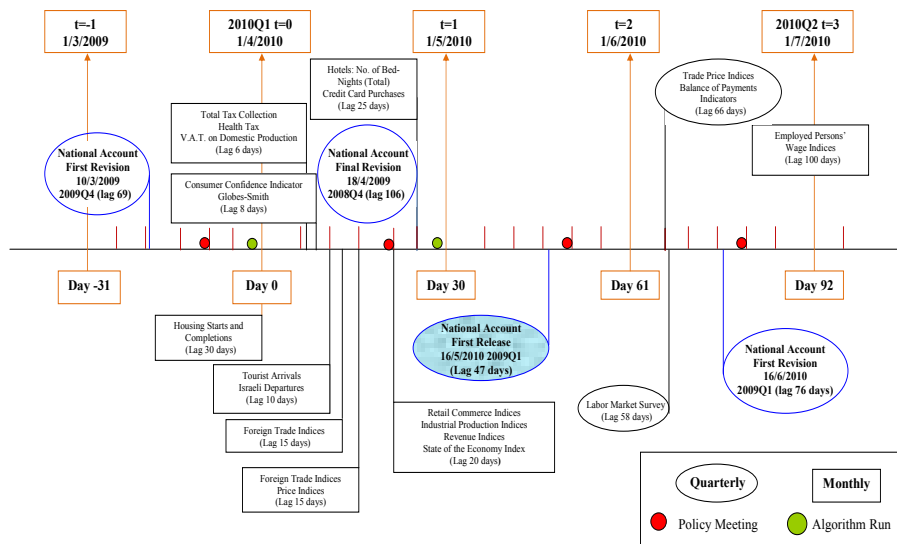
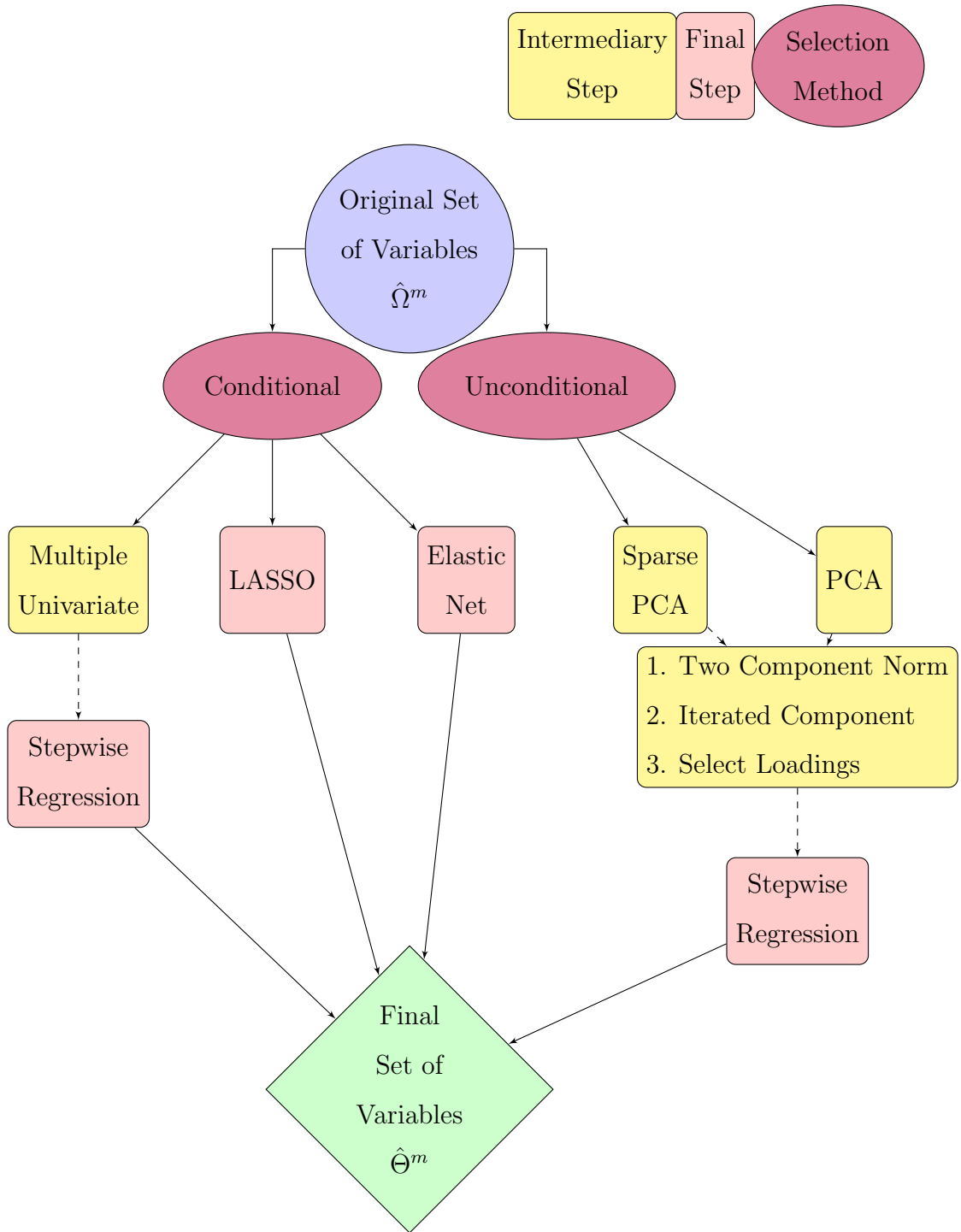


Figure 9: Flowchart of Methods Applied



## A Tables

Table 1: Probability of Variables to Enter the Final Subset

<b>Table 1</b>	
<b>Conditional Methods</b>	
<b>Univariate Multiple Regression</b>	
Prob Chosen	Description
92%	TMB :motors, electronic components
32%	PMI:Production :output component
28%	BOI Companies Survey : Communication
24%	US Exports : Goods
24%	TPR :electronic components and different industrial equip.
16%	PMI:Domestic :supply :time component
4%	Cushing Crude Oil Price
4%	US Exports: Serivces
4%	PMI:Stock :purchase :quantity component
4%	Employer Survey : Transportation
<b>Elastic Net</b>	
<b>Alpha=0.4</b>	
<b>Quantile=0.75</b>	
Prob Chosen	Description
100%	High Tech Stock Index
100%	Cushing Crude Oil Price
100%	US Exports : Goods
100%	PMI : Employment :level component
Continued on next page	

<b>Table 1 – continued from previous page</b>	
100%	Tourist bed-nights in authorized and unauthorized hotels
100%	TPR :textiles
100%	TPR :electronic motors, components and transport equip.
100%	TPR :electronic components and different industrial equip.
100%	TMB :motors, electronic components
96%	Globes-Smith CCI
<b>Quantile=0.50</b>	
Prob Chosen	Description
100%	US Exports : Goods
100%	PMI : Employment :level component
100%	TPR :electronic motors, components and transport equip.
100%	TPR :electronic components and different industrial equip.
100%	TMB :motors, electronic components
96%	Cushing Crude Oil Price
96%	TMB :electronic components and different industrial equipment
92%	High Tech Stock Index
92%	PMI : Euro 16
80%	General Stock index
<b>Quantile=0.25</b>	
Prob Chosen	Description
100%	US Exports : Goods
100%	PMI : Employment :level component
100%	PMI:Stock :purchase :quantity component
100%	TPR :electronic components and different industrial equip.
100%	TMB :motors, electronic components
Continued on next page	

<b>Table 1 – continued from previous page</b>	
100%	TMB :electronic components
100%	TMB :electronic components and different industrial equipment
96%	PMI : Euro 16
92%	TPR :electronic motors, components and transport equip.
84%	BOI Companies Survey : Communication
<b>LASSO</b>	
<b>Quantile=0.75</b>	
Prob Chosen	Description
100%	Tourist bed-nights in authorized and unauthorized hotels
100%	TPR :textiles
100%	TMB :Machinery and equipment
100%	TMB :motors, electronic components
96%	Cushing Crude Oil Price
92%	High Tech Stock Index
92%	US Exports : Goods
92%	TPR :electronic components and different industrial equip.
84%	RT : Consumption goods :Other
84%	Health Tax
<b>Quantile=0.50</b>	
Prob Chosen	Description
100%	TPR :electronic components and different industrial equip.
100%	TMB :motors, electronic components
92%	Cushing Crude Oil Price
92%	US Exports : Goods
68%	PMI : Employment :level component
Continued on next page	

<b>Table 1 – continued from previous page</b>	
68%	PMI : Euro 16
60%	Industrial Firms Stock Index
60%	Health Tax
60%	Tourist bed-nights in authorized and unauthorized hotels
60%	TPR :electronics communication equipment
<b>Quantile=0.25</b>	
Prob Chosen	Description
100%	TMB :motors, electronic components
84%	US Exports : Goods
84%	TPR :electronic components and different industrial equip.
80%	PMI:Stock :purchase :quantity component
72%	PMI : Euro 16
60%	BOI Companies Survey : Communication
36%	PMI : Employment :level component
36%	PMI:Production :output component
36%	TMB :electronic components
16%	PMI:Domestic :supply :time component
<b>Unconditional Methods</b>	
<b>PCA</b>	
<b>Two Component Norm Selection</b>	
Prob Chosen	Description
28%	Industrial Firms Stock Index
20%	PMI:Domestic :supply :time component
20%	TMB :Jewelry, goldsmiths' and silversmiths' articles
16%	Employer Survey : Education
Continued on next page	



<b>Table 1 – continued from previous page</b>	
8%	General Stock index
8%	Israel Exports :Services (NIS)
4%	High Tech Stock Index
4%	TMB :food products
4%	TMB :food products, beverages and tobacco products
4%	TMB :textiles
<b>Iterated Component Selection</b>	
Prob Chosen	Description
44%	PMI:Domestic :supply :time component
32%	TMB :Jewelry, goldsmiths' and silversmiths' articles
32%	Employer Survey : Building
28%	Cushing Crude Oil Price
28%	Manufacturing exports (NIS)
24%	Industrial Firms Stock Index
20%	DOLLAR/NIS EXCHANGE RATE
16%	Health Tax
16%	Israel Exports :Services (NIS)
16%	PMI:Domestic :orders component
<b>SPCA</b>	
<b>Two Component Norm Selection</b>	
Prob Chosen	Description
16%	Triple trade index : Nominal
16%	PMI : Employment :level component
16%	Employer Survey : Food
12%	PMI:Domestic :supply :time component
Continued on next page	

<b>Table 1 – continued from previous page</b>	
12%	TPR :electronic motors, components and transport equip.
12%	TMB :motors, electronic components
12%	PMI : Euro 16
8%	US Exports: Services
8%	Revenue index :Community, social, personal and other services
8%	Revenue index :Commerce and services
Iterated Component Selection	
<b>Prob Chosen</b>	<b>Description</b>
32%	Tel Aviv 100
28%	Cushing Crude Oil Price
28%	TMB :industrial equipment for control and supervision
24%	PMI:Domestic :supply :time component
20%	Hotels :no. of bed-nights in tourist hotels :Israeli
20%	TMB :wearing apparel
16%	General Stock index
12%	US Exports: Services
12%	Israel Exports :Services (NIS)
12%	PMI:Stock of finished :goods component

Table 2: List of Variables in the General Set

Index	Description
	BOI: Bank of Israel, RT: Retail Trade, TPR: Industrial Production Index, TMB: Employed Persons' Index, PMI: Purchasing Manager's Index, CCI: Consumer Confidence Index, RT: Retail Trade, THP: Man-Hours Worked Index
1	BOI Companies Survey : Building
2	BOI Companies Survey : Communication
3	BOI Companies Survey : Industry
4	BOI Companies Survey : Retail
5	Capital Utilization Index
6	Cushing Crude Oil Price
7	Dollar/NIS Exchange Rate
8	Employer Survey : Agriculture
9	Employer Survey : Building
10	Employer Survey : Education
11	Employer Survey : Financial
12	Employer Survey : Food
13	Employer Survey : Health
14	Employer Survey : Industry
15	Employer Survey : Real
16	Employer Survey : Trade
17	Employer Survey : Transportation
18	EURO/NIS EXCHANGE RATE
19	EURO/NIS EXCHANGE RATE
20	General Concert Bonds Stock index

Continued on next page

Table 2 – continued from previous page

Index	Description
	BOI: Bank of Israel, RT: Retail Trade, TPR: Industrial Production Index, TMB: Employed Persons' Index, PMI: Purchasing Manager's Index, CCI: Consumer Confidence Index, RT: Retail Trade, TPH: Man-Hours Worked Index
21	General Stock index
22	Globes-Smith CCI
23	GOLD : Market Rate
24	GOLD : Market Rate
25	Gross Capital Stock : Business Sector
26	Health Tax
27	High Tech Stock Index
28	Hotels : No. of bed-nights in tourist hotels :total
29	Hotels : No. of bed-nights in tourist hotels :Israeli
30	Housing completions :Total
31	Housing starts :public sector
32	Housing starts :Total
33	Imports :consumer goods (NIS)
34	Imports :Investment goods (NIS)
35	Imports :Net (NIS)
36	Industrial Firms Stock Index
37	Israel Exports :Goods (NIS)
38	Israel Exports :Services (NIS)
39	Israel Imports :Goods (NIS)
40	Israel Imports :Services (NIS)
41	Manufacturing exports (NIS)
Continued on next page	

Table 2 – continued from previous page

Index	Description
	BOI: Bank of Israel, RT: Retail Trade, TPR: Industrial Production Index, TMB: Employed Persons' Index, PMI: Purchasing Manager's Index, CCI: Consumer Confidence Index, RT: Retail Trade, TPH: Man-Hours Worked Index
42	Michigan CCI
43	MSCI : Currency(NIS)
44	No. of tourist arrivals :total
45	No. of tourist arrivals, by air passengers
46	PMI : Employment :level component
47	PMI : Euro 16
48	PMI : USA
49	PMI:Domestic :orders component
50	PMI:Domestic :supply :time component
51	PMI:Global :orders component
52	PMI:Import :supply :time component
53	PMI:Production :output component
54	PMI:Raw :material :stock :levels component
55	PMI:Stock :purchase :prices component
56	PMI:Stock :purchase :quantity component
57	PMI:Stock of finished :goods component
58	Price index of dwellings
59	Purchasing Managers Index :Total
60	Real Effective Exchange Rate
61	Residential building :Completions :private sector
62	Residential building :Completions :public sector
Continued on next page	

Table 2 – continued from previous page

Index	Description
	BOI: Bank of Israel, RT: Retail Trade, TPR: Industrial Production Index, TMB: Employed Persons' Index, PMI: Purchasing Manager's Index, CCI: Consumer Confidence Index, RT: Retail Trade, TPH: Man-Hours Worked Index
63	Residential building :Starts :private sector
64	Revenue index :Accommodation services and restaurants
65	Revenue index :Banking, insurance and other Financial institutions
66	Revenue index :Business activities
67	Revenue index :Commerce and services
68	Revenue index :Community, social, personal and other services
69	Revenue index :Education
70	Revenue index :Health, welfare & social work services
71	Revenue index :Wholesale and retail trade, and repairs
72	RT : Consumption goods :Other
73	RT : Durables
74	RT : Footwear
75	RT : Textile and clothing
76	RT : Total excl. gas, fertilizers and petroleum
77	RT :Food
78	RT :Kitchen and house accessories
79	RT :Petroleum
80	Tel Aviv 100
81	THP : basic metal
82	THP : beverages and tobacco products
83	THP : chemicals and their products
Continued on next page	

**Table 2 – continued from previous page**

Index	Description
	BOI: Bank of Israel, RT: Retail Trade, TPR: Industrial Production Index, TMB: Employed Persons' Index, PMI: Purchasing Manager's Index, CCI: Consumer Confidence Index, RT: Retail Trade, TPH: Man-Hours Worked Index
84	THP : components
85	THP : electronics communication equipment
86	THP : furniture
87	THP : High technology
88	THP : industrial equip. for control
89	THP : industry : index
90	THP : Jewelry, goldsmiths' and silversmiths' articles
91	THP : Low technology
92	THP : Machinery and equipment
93	THP : Medium-high technology
94	THP : Medium-low technology
95	THP : metal products
96	THP : motors, electronic components and equip.
97	THP : other mining and quarrying
98	THP : textiles & wearing apparel
99	THP : textiles
100	THP : Transport equipment
101	THP : wood and its products & furniture
102	TMB : chemicals and their products
103	TMB :basic metal
104	TMB :beverages and tobacco products
Continued on next page	

**Table 2 – continued from previous page**

Index	Description
	BOI: Bank of Israel, RT: Retail Trade, TPR: Industrial Production Index, TMB: Employed Persons' Index, PMI: Purchasing Manager's Index, CCI: Consumer Confidence Index, RT: Retail Trade, TPH: Man-Hours Worked Index
105	TMB :electric motors and electric distribution apparatus
106	TMB :electronic components
107	TMB :electronic components and different industrial equipment
108	TMB :food products
109	TMB :food products, beverages and tobacco products
110	TMB :footwear, leather and its products
111	TMB :furniture
112	TMB :industrial equipment for control and supervision
113	TMB :Jewelry, goldsmith's and silversmith's articles
114	TMB :Machinery and equipment
115	TMB :Manufacture of plastic and rubber products
116	TMB :manufacturing n.e.c
117	TMB :metal products
118	TMB :motors, electronic components
119	TMB :non-metallic mineral products
120	TMB :other mining and quarrying
121	TMB :paper and its products
122	TMB :publishing and printing
123	TMB :textiles
124	TMB :textiles & wearing apparel
125	TMB :Transport equipment
Continued on next page	



**Table 2 – continued from previous page**

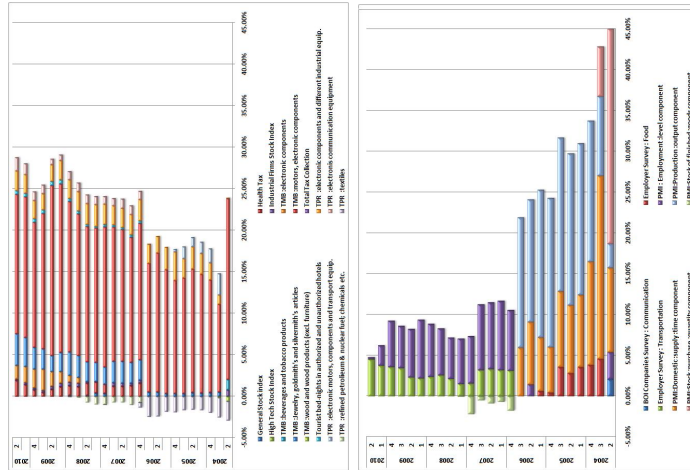
Index	Description
	BOI: Bank of Israel, RT: Retail Trade, TPR: Industrial Production Index, TMB: Employed Persons' Index, PMI: Purchasing Manager's Index, CCI: Consumer Confidence Index, RT: Retail Trade, TPH: Man-Hours Worked Index
126	TMB :wearing apparel
127	TMB :wood and its products & furniture
128	TMB :wood and wood products (excl. furniture)
129	Total Tax Collection
130	Tourist bed-nights in authorized and unauthorized hotels
131	TPR : Other Branches
132	TPR :beverages and tobacco products
133	TPR :building products :paper and its products
134	TPR :electronic components and different industrial equip.
135	TPR :electronic motors, components and transport equip.
136	TPR :electronics communication equipment
137	TPR :food products, beverages and tobacco products
138	TPR :furniture
139	TPR :High technology
140	TPR :industrial equipment for control and supervision
141	TPR :Jewelry, goldsmith's and silversmith's articles
142	TPR :Low technology
143	TPR :Medium-high technology
144	TPR :Medium-low technology
145	TPR :other mining and quarrying
146	TPR :refined petroleum & nuclear fuel; chemicals etc.
Continued on next page	

**Table 2 – continued from previous page**

<b>Index</b>	<b>Description</b>
	BOI: Bank of Israel, RT: Retail Trade, TPR: Industrial Production Index, TMB: Employed Persons' Index, PMI: Purchasing Manager's Index, CCI: Consumer Confidence Index, RT: Retail Trade, TPH: Man-Hours Worked Index
147	TPR :textiles
148	TPR :textiles & wearing apparel & footwear etc.
149	TPR :total (excl. diamonds)
150	TPR :Transport equipment
151	TPR :wood and its products & furniture
152	TPR :Metal and machinery
153	Treasury bills: Fixed interest 1 month to redemption
154	Triple trade index : Nominal
155	Triple trade index : Real
156	US Exports : Goods
157	US Exports: Services
158	V.A.T. on Domestic Production

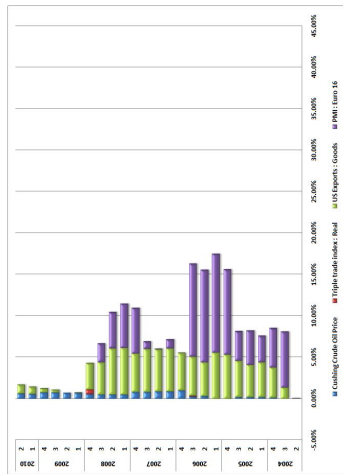
# B Figures

Figure 1: Size of Coefficients in each Period (LASSO)



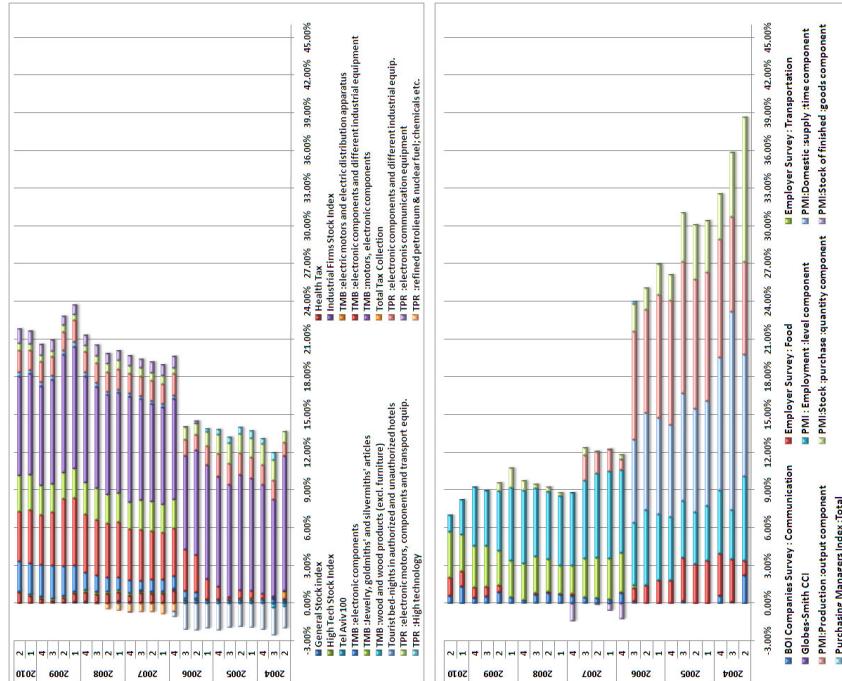
(a) Domestic

(b) Market Expectations



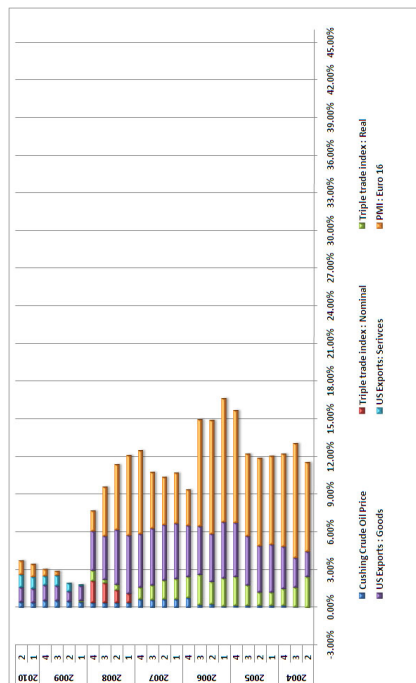
(c) Global

Figure 2: Size of Coefficients in each Period (Elastic Net  $\alpha = 0.4$ )



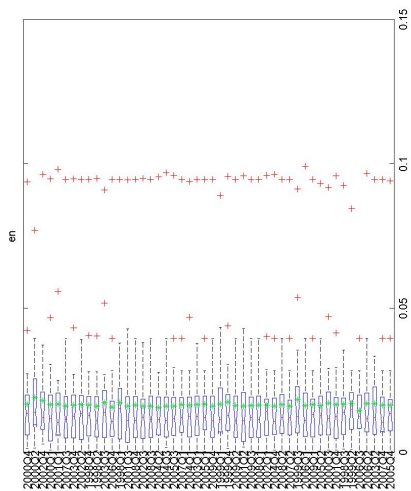
(a) Domestic

(b) Market Expectations

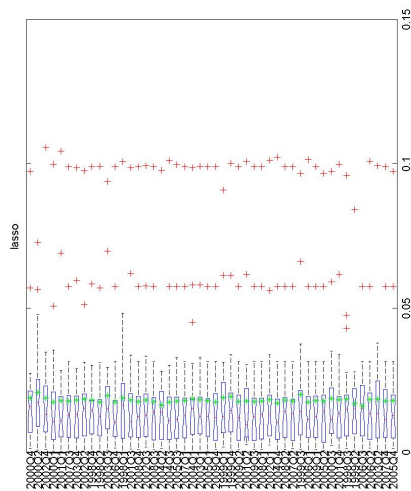


(c) Global

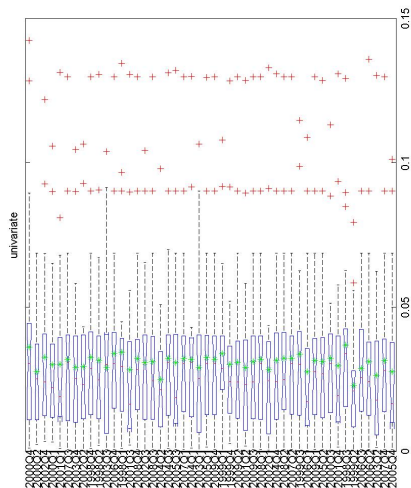
Figure 3: Jackknife Boxplots of Conditional Methods



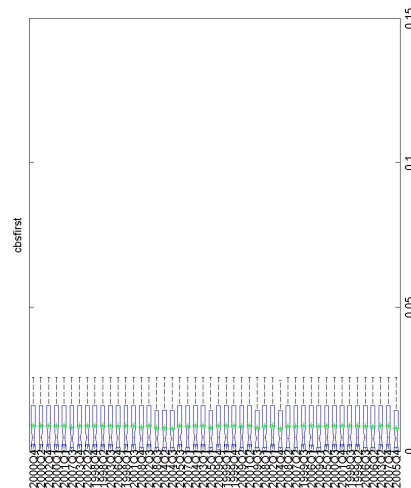
(a) Elastic Net



(b) LASSO

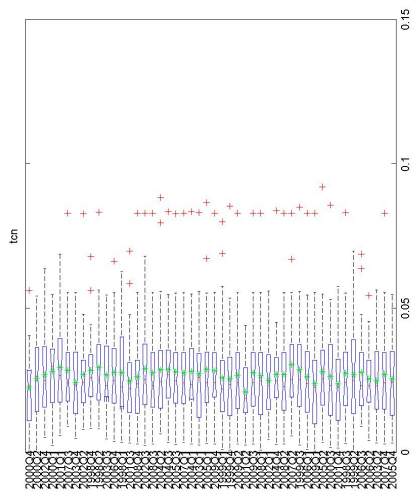


(c) Univariate

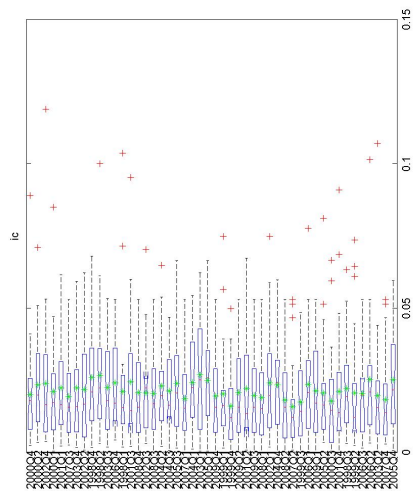


(d) CBS<sub>First</sub>

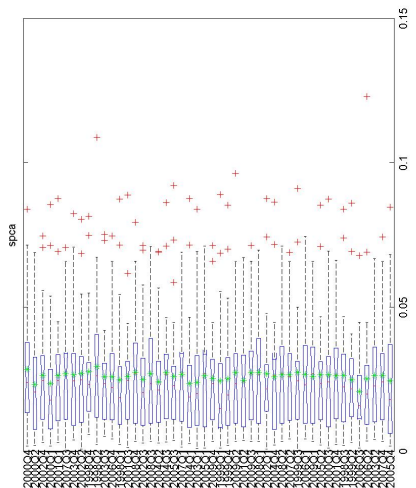
Figure 4: Jackknife Boxplots of Unconditional Methods



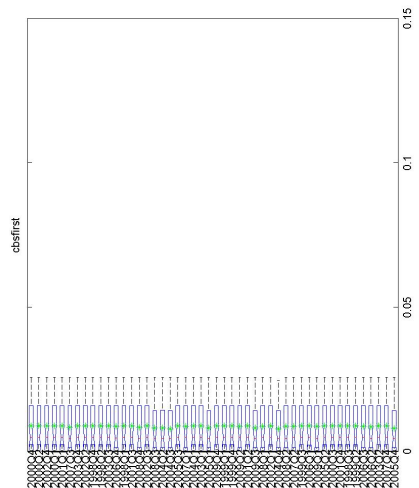
(a) TCN



(b) IC



(c) PCA



(d) CBS<sub>First</sub>

### 3. ESTIMATING THE UNCERTAINTY OF DATA REVISIONS IN ASYNCHRONOUS STATIONARY PROCESSES USING HIDDEN MARKOV MODELS

**Status:** This chapter was submitted to a peer-review journal and is currently under review.

Estimating the Uncertainty of Data Revisions  
in Asynchronous Stationary Processes using  
Hidden Markov Models

Jonathan Sidi<sup>1</sup>

<sup>1</sup>Department of Statistics, Hebrew University of Jerusalem



### **Abstract**

A typical time series can be defined as a stochastic process that accumulates new data at fixed time intervals. Once these realizations are part of the observed process they are treated as constant. There are cases in which the stochastic process may not be directly observed. This necessitates their estimation using sequentially accumulated samples that provide new information relevant to the period of interest. As a result, after the initial estimation all points in the stochastic process are repeatedly revised.

We propose in this paper a method to estimate the uncertainty found in the revision process. We use this estimation to generate an asymmetrical prediction interval of a subsequent revision to the currently published activity period. The interval is a function of the maturity of each time period within the current vintage of the growth process. Our approach relaxes a common assumption that the revision process is a homogeneous one and not a mixture of a number of different processes. We postulate that the revision process is a function of the state of the growth process, thereby creating the necessity to model the process state through hidden Markov models and estimate model parameters relating to each one.

The methodology is tested on historical data of the Israel Gross Domestic Product (GDP). We find that there is a significant difference in the size and sign of revisions dependent on the state of the growth rate. More specifically, when the initial publication is in a low growth period, the growth is overestimated and subsequent revisions lower the growth rate, and conversely when the initial publication is in a high growth period the growth is underestimated and the subsequent revisions increase the growth rate. Out of sample tests are conducted and show that the prediction interval estimate of the initial publication

correctly identifies both the future first and third revisions at a rate of 70% within an interval width of 0.35% annualized growth.

## 1 Introduction

A typical time series can be defined as a stochastic process that accumulates new data at fixed time intervals. Once these realizations are part of the observed process they are treated as constant. There are cases in which the stochastic process may not be directly observed. This necessitates their estimation using sequentially accumulated samples that provide new information relevant to the period of interest. As a result, after the initial estimation all points in the stochastic process are repeatedly revised. We model the revision process using a hidden Markov model to capture this inherent uncertainty.

We characterize the process as asynchronous, in that random variables in the sequence are revised at each time period, whereas new random variables are added to the sequence at a lower frequency. This creates two levels of uncertainty in the process: the uncertainty between the random variables in the process and uncertainty pertaining to a given random variable which is dependent on the state of the overall process and its maturity since initial estimation.

Examples can be found in the publication of official statistics, financial reports, banks' internal credit ratings and the estimation of fetal growth. Indicators relating to the state of these processes are unobserved and must be modeled using proxy information. In such cases the true value of any aggregate data is unobserved, and therefore the measurement error of revisions is likewise unobserved. Such activity indicators can be viewed as a series of independent random surveys with respect to a period of activity, Kapetanios & Yates (2004).

In the field of official statistics, every three months a new period of activity is estimated and after which is revised once a month utilizing new information. Financial reports are published every three months by publicly

---

traded companies. The initial publication contains a number of estimated indicators which represent the current fiscal state of the company. At the time of the next period publication revised estimates are published regarding previously reported periods, which contain information that was not available at the previous period. When estimating fetal growth a number of observable measurements, such as femur length and head circumference, are taken during the 11th week and then every four-six weeks until birth. These estimates are aggregated to create the fetal growth, which has a high rate of measurement error. In all examples the uncertainty found in the early publications can have great implications regarding the action each practitioner takes. The monetary policy of central banks through the official statistics, investment actions of the market participants based on the published financial reports and physician interventions to counteract faltering fetal growth.

We propose in this paper a method to estimate the revision process uncertainty. We use this estimation to generate an asymmetrical prediction interval of an upcoming revision of a currently published activity period. These intervals are a function of the maturity of each time period within the current vintage of the growth process. Contemporary literature, such as Cunningham et al. (2012), Anderson & Gascon (2009) and Jacobs & Van Norden (2011), model revisions of economic growth base their models on two major assumptions. An explicit assumption is that the published growth can be de-constructed into three parts: the true growth, a time invariant publication bias of a given maturity, and the serially correlated measurement error associated with the publication maturity. The decay rate is estimated over the full sample and is not a function of the level of maturity. We find that empirically the revisions do decay over a long horizon of nearly ten years, these are comparable to the revision decay at of the US and England GDPs.

Thus, we do not find it pertinent to model the full path of revisions due to its negligible use for real time policy decision making. An implicit assumption is made in that the revision process is a homogeneous one and not a mixture of a number of different processes. We postulate that the revision process is a function of the state of the growth process, thereby creating the necessity to model the growth state and estimate model parameters relating to each one. An unsupervised learning algorithm is applied to classify different revision behavior present in past vintages. We estimate the transition probability of the process growth being in a given state of growth through a Markov switching model.

The application of these prediction intervals gives the policy maker a level of confidence with respect to the probability the estimate they are given initially will be revised and pass predetermined decision thresholds. Instead of basing this on either endogenous information currently at hand or inherent model error that can be modulated, we propose to use empirical revision distributions that are exogenous to the growth process estimation framework.

The following sections of the paper will present the formal definition of the process, the hidden Markov model, the derivation of the prediction interval and finally a case study using the GDP publications from Israel.

## 2 Growth Process

Let  $\{y_t^k\}_{t=1, k=0}^{T, K}$ ,  $t \in \mathbb{N}^+$ ,  $k \in \mathbb{N}^0$  be a discrete time, continuous space, stochastic and stationary process. Where  $t$  denotes the time period which the random variable is initially estimated. Thereafter, at each time period, each random variable is revised, i.e. re-estimated, where  $k$  denotes the maturity of revision. Given a period of publication  $t$ , the number of revisions that have been

been applied to a random variable is defined as  $K(n, t) = n(t - 1) - I_{\{t > 1\}}$ , where  $n$  is the ratio between the frequency of  $t$  and the frequency of revision and  $I$  is an indicator function. For simplicity of notation going further  $K \equiv K(n, t)$ . This framework generates different length processes dependent on period of initial estimation. A stylized representation of this process is presented in Table 1 under the assumption that  $n=4$ :

$$y_t^k = \begin{bmatrix} y_1^0 & y_1^1 & y_1^2 & y_1^3 & y_1^4 & y_1^5 & y_1^6 & y_1^7 & y_1^8 & \dots \\ & & & y_2^0 & y_2^1 & y_2^2 & y_2^3 & y_2^4 & y_2^5 & \dots \\ & & & & & & y_3^0 & y_3^1 & y_3^2 & \dots \end{bmatrix} \quad (1)$$

We assume that  $y_t^k$  is a first order autoregressive process, dependent on both vintage and activity period. In the standard AR(1) model we would condition on the previous period, but in this case we have more information regarding the previous period. When the current activity period is initially realized, the previous one has been revised  $K$  times. We can use its historical revisions which contain information regarding  $t - 1$ . We add an additional covariate to the model  $y_{t-1}^{k+m}$  which represents  $y_{t-1}^k$  at revision maturity  $m$ . This revision maturity parameter is used to create flexibility of which type of additional information is used in the model. As seen in the example above  $y_3^0$  can be conditioned on four random variables from the previous period,  $y_2^0, y_2^1, y_2^2, y_2^3$ , where the earliest one is in fact four higher frequency time periods prior to  $y_3^0$ . In practice this maturity represents a large horizon in the high frequency time line, where the time difference between activity period  $t$  and  $t - 1$  increases at a fixed ratio. We incorporate an additional prior vintage of  $y_{t-1}$  an exogenous variable to the standard AR(1) model, thus constructing an ARIMAX model, Hamilton (1990).

$$y_t^k = a_0 + a_1 y_{t-1}^k + a_2 y_{t-1}^{k+m} + \epsilon_t, \quad \epsilon_t \stackrel{iid}{\sim} N(0, \sigma_a^2) \quad (2)$$

Using these two covariates we can separate the effect ‘new information’, added to the previous period ( $t - 1$ ), has on the current period ( $t$ ), which is being predicted. The model parameters are denoted in a vector  $\alpha = (a_0, a_1, a_2, \sigma_a^2)'$ .

We find empirical evidence which shows that the revision process within a given vintage is a function of the state of the overall growth process. The GDP growth process in many economies has been found to follow three distinct growth regimes, low/mid/high, where the low growth includes also deceleration, i.e. negative growth. We show that the revision process of the Israel GDP differs within each growth regime. The low growth regimes have heavy lower tail revision distributions and the high growth regimes have heavy upper tail revision distributions. Our objective is to estimate the probability of which regime the rate of growth is at in each activity period of a vintage.

We model 2 as a Markov switching model, in which there are a total of  $N$  regimes. Let  $s_t^k$  be an unobserved scalar random variable which takes on integer values  $j \in 1 \dots N$ . Given  $s_t^k = j$  we define the model as:

$$y_t^k = a_{0j} + a_{1j} y_{t-1}^k + a_{2j} y_{t-1}^{k+m} + \epsilon_t, \quad \epsilon_t \stackrel{iid}{\sim} N(0, \sigma_{aj}^2) \quad (3)$$

Given constant  $k$  and  $m$  the  $T \times 1$  vector  $y_t^k$  depends only on the current and previous state variables and the previous activity period corresponding publication vintage, the historical vintage  $s_t^k, s_{t-1}^k, y_{t-1}^k, y_{t-1}^{k+m}$ , and on a vector of parameters  $\theta$ . The parameter vector is defined as a stacked vector of

$\theta = (\alpha'_1 \dots \alpha'_N)'$  where the vector  $\alpha_j = (a_{0j}, a_{1j}, a_{2j}, \sigma_{aj}^2)'$  corresponds to the model parameters of  $s_t^k = j$ . Formally the Markovian process is defined as

$$\begin{aligned} f(y_t^k | s_t^k, s_{t-1}^k, \dots, y_{t-1}^k, y_{t-2}^k, \dots, y_{t-1}^{k+m}, y_{t-2}^{k+m}, \dots; \theta) \\ = f(y_t^k | s_t^k, y_{t-1}^k, y_{t-1}^{k+m}; \theta) \\ = f(y_t^k | s_t^k, z_{t-1}^{k,m}; \theta), \end{aligned} \quad (4)$$

where  $z_{t-1}^{k,m} = ((y_{t-1}^k)', (y_{t-1}^{k+m})')'$ . For each time period the conditional density is defined as a function of a given regime:

$$\eta_{jt} = f(y_t^k | s_t^k = j, z_{t-1}^{k,m}; \theta) \sim N \left( \left( z_{t-1}^{k,m} \right)' \beta_j, \sigma_j^2 \right), \quad (5)$$

Where  $\beta_j$  is the coefficient vector for the  $j^{\text{th}}$  regression. There are thus  $N$  conditional densities which are collected in an  $(N \times 1)$  vector  $\eta_t = (\eta'_{1t}, \dots, \eta'_{Nt})'$ . We therefore need to make an inference regarding the state value of  $s_t^k$  given all the information we know with certainty up to that time.

## 2.1 Hamilton Filter

For each time period  $t = 1 \dots T$  we estimate the probability to be in each value of  $s_t^k$

$$\xi_{jt}^k = \Pr(s_t^k = j | z_t^{k,m}; \theta), \quad (6)$$

where for each time period the probabilities sum to unity. The inference regarding the probability to be in a given state is performed iteratively for each  $t$  according to



$$\hat{\xi}_{jt}^k = \frac{f(y_t^k, s_t^k = j | z_{t-1}^{k,m}; \theta)}{f(y_t^k | z_{t-1}^{k,m}; \theta)}, \quad (7)$$

where the conditional joint density  $f(y_t^k, s_t^k = j | z_{t-1}^{k,m}; \theta)$  is given by

$$\begin{aligned} f(y_t^k, s_t^k = j | z_{t-1}^{k,m}; \theta) &= \sum_{i=1}^N \Pr(s_t^k = j | s_{t-1}^k = i) \Pr(s_{t-1}^k = i | z_{t-1}^{k,m}; \theta) \eta_{jt} \\ &= \sum_{i=1}^N p_{ij}^k \hat{\xi}_{i,t-1}^k \eta_{jt}. \end{aligned} \quad (8)$$

It is assumed that the  $s_t^k$  evolves according to a Markov chain that is independent of the current or past observations of  $z_t^{k,m}$ :

$$p_{ij}^k \equiv \Pr(s_t^k = j | s_{t-1}^k = i, s_{t-2}^k = i', \dots, z_t^{k,m}, z_{t-1}^{k,m}) = \Pr(s_t^k = j | s_{t-1}^k = i). \quad (9)$$

Using 8 and 9 we can estimate the conditional distribution of  $s_t^k$

$$\hat{\xi}_{jt}^k = \frac{f(y_t^k, s_t^k = j | z_{t-1}^{k,m}; \theta)}{f(y_t^k | z_{t-1}^{k,m}; \theta)} = \frac{\sum_{i=1}^N p_{ij}^k \hat{\xi}_{i,t-1}^k \eta_{jt}}{\sum_{j=1}^N \sum_{i=1}^N p_{ij}^k \hat{\xi}_{i,t-1}^k \eta_{jt}}, \quad (10)$$

where the conditional density of the  $t^{\text{th}}$  observation is the sum of the  $N$  modalities for each  $j$ . Finally, for each iteration we evaluate the sample conditional log likelihood of the observed data:

$$\log f(y_1^k \dots y_T^k | y_0^k) = \sum_{t=1}^T \log f(y_t^k | z_{t-1}^{k,m}; \theta) \quad (11)$$

Smoothed inferences are calculated using Kim's algorithm, Kim (1994), starting from  $\xi_{jT}$  obtained from equation 10 we iterate backwards  $t = T - 1, T - 2, \dots$

$$\hat{\xi}_{t|T}^k = \hat{\xi}_{t|t}^k \odot \{(P^k)' [\hat{\xi}_{t+1|T}^k (\div) \hat{\xi}_{t+1|t}^k]\}, \quad (12)$$

where  $\hat{\xi}_{t|t}^k$ , represents a stacked  $N \times 1$  vector whose  $j^{th}$  element is  $\hat{\xi}_{jt}^k$  and  $(\div)$ ,  $\odot$  denote element-by-element division and multiplication respectively. We update the transition probability using the smoothed inferences through

$$\hat{p}_{ij}^k = \frac{\sum_{t=2}^T \Pr(s_t^k = j, s_{t-1}^k = i; \hat{\theta})}{\sum_{t=2}^T \Pr(s_{t-1}^k = i; \hat{\theta})}, \quad (13)$$

where  $\hat{\theta}$  denotes the vector of updated maximum likelihood estimates.

The EM algorithm is initialized with an initial guess to the parameters  $\theta^{(0)}$ , the initial state probabilities  $\xi_{0j} = N^{-1} \forall j$ , and a uniform probability transition matrix  $p_{ij}^k = N^{-1}, \forall i, j$ . Using these initial values we evaluate the filtered and smoothed probabilities and update the transition matrix (E-step). Using the estimated probabilities the parameter estimates are re-estimated to generate  $\theta^{(1)}$  and the model likelihood is updated (M-step). These steps are iterated in the same fashion to generate  $\theta^{(2)}, \theta^{(3)}, \dots$  until convergence is achieved. We set the convergence criterion to be  $\|\theta^{\text{new}} - \theta^{\text{old}}\| \leq 10^{-8}$ . This is applied to different starting points to avoid local solutions.

## 2.2 Model Inference

Model 3 allows us to infer which growth regime the process is most likely in and the transition probability to be in a given state for the next time period. If we were to forecast the m-periods ahead under model 3 we would in fact be forecasting the probability to be in a growth regime in future activity periods. This however is not the goal of this paper, we are interested in generating prediction intervals for the revisions to the current vintage published.

An intuitive example to better understand the difference is the re-evaluation of interest rates in a central bank. When the gross domestic product is initially published policy makers need to know the level of uncertainty regarding the subsequent revisions to the new activity period. If we were to forecast model 3 we would be supplying them information regarding the expected regime of the next activity period. There are many models that supply such predictions such as the Dynamic Stochastic General Equilibrium structural models, for examples see Romer (2012).

### 3 Non-Parametric Prediction Intervals

Given an updated estimation for a given activity period and the distribution of realized revisions we estimate the interval that the subsequent revisions will most likely be in, given a predetermined probability. This in contrast to confidence intervals or credible intervals, that are prevalent in frequentist and Bayesian inference. In those inferences the objective is predict the distribution of parameter estimates to the quantity of interest that cannot be observed.

We first characterize the revision process that are the basis of the prediction intervals. We denote the revision process of  $y_t^k$  as

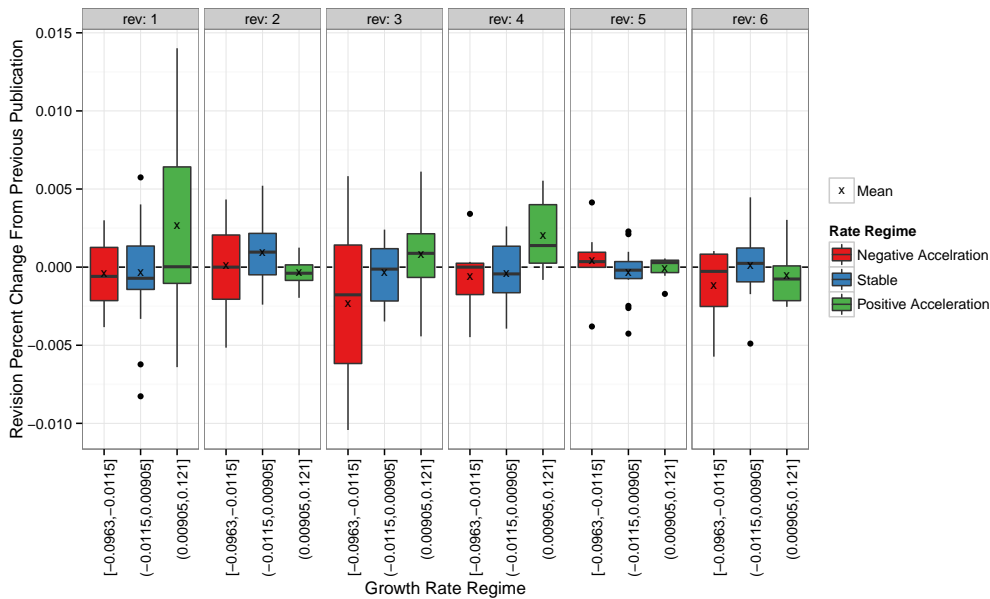
$$x_t^k = \frac{y_t^k - y_t^{k-1}}{y_t^k}, k = 1 \dots K. \quad (14)$$

Within a given activity period  $t$  we assume that the revision process is a function of the unobserved state variable  $s_t^{k+m}$ , which corresponds to the relevant horizon  $m$ . For example in Table 1 the upper row consists of the initial estimate  $y_1^0$  and eight subsequent revisions,  $y_1^1 \dots y_1^8$ . For this time period we would construct the revision process  $x_1^1 \dots x_1^8$ , which itself is a

random process. After constructing such variables for each time period, we combine them into the matrix  $X = (x_1, x_2, \dots, x_T)'$ . The first column of  $X$  gives the distribution of first revisions for all time periods, the second gives the distribution of the subsequent revision and so on.

In the case study in the following section it is shown that the revision distributions of the Israel GDP are a mixture of three distinct processes which are characterized by the unobserved state variable we modeled in the previous sections.

Figure 1: Revision Distribution by Regime  
Sample Period 2005m9-2014m3



A depiction of the revision distributions are shown in Figure 1. Due to the high levels of skewness present in the revision distributions, our objective is to construct prediction intervals with respect to their empirical distribution.

### 3.1 Construction of Prediction Intervals

To construct the prediction intervals we evaluate the expected revision at the upper and lower percentiles of the revision distribution of each maturity conditioned on the estimated growth regime state.

$$X_l^k = 100 \times (1 - p^{th}) \text{ percentile of } X^k \quad (15)$$

$$X_u^k = 100 \times p^{th} \text{ percentile of } X^k \quad (16)$$

The current process growth  $y_t^k$  is updated with the weighted sum of the revision distribution at the lower and upper bounds, multiplied by the probability to be in each state at time  $t$ ,  $\xi_{jt}^k$ .

$$y_t^{k+m,l} = (1 + y_t^k) \left[ 1 + \sum_{j=1}^N X_l^k \xi_{jt}^k \right] \quad (17)$$

$$y_t^{k+m,u} = (1 + y_t^k) \left[ 1 + \sum_{j=1}^N X_u^k \xi_{jt}^k \right]. \quad (18)$$

While the interval is symmetric in percentiles, it is not symmetric in the terms of values of the process growth rate. Thus,  $y_t^{k+m,l}$  and  $y_t^{k+m,u}$  are endpoints of the  $100p\%$  prediction interval for  $y^{k+m}$

$$\Pr \left( y_t^{k+m,l} < y_t^{k+m} < y_t^{k+m,u} \right) = p. \quad (19)$$

For example observing the realizations of  $y_t^0$  and  $y_{t-1}^3$  we want to evaluate the revised values of the rate of process growth within a band that covers 80% of the empirical distribution. We define the lower bound as the 10<sup>th</sup> percentile and the upper bound as the 90<sup>th</sup> percentile of  $y_t^1$ ,

$$y_t^{1,l} = (1 + y_t^0) \left[ 1 + \sum_{j=1}^N X_l^1 \Pr(s_t^0 = j | s_{t-1}^0 = i, y_{t-1}^3) \right] \quad (20)$$

$$y_t^{1,u} = (1 + y_t^0) \left[ 1 + \sum_{j=1}^N X_u^1 \Pr(s_t^0 = j | s_{t-1}^0 = i, y_{t-1}^3) \right]. \quad (21)$$

## 4 Model Initialization

A classification algorithm is defined to categorize the revision distribution conditional on features of the stationary the process. An unsupervised learning scheme is used to label each data point to its nearest distribution, through a non-parametric hypothesis test. We use a non-parametric test for two reasons. The first is that the revision distributions have high levels of skewness that changes throughout the maturity horizon, thus modeling such a process would add many assumptions and create a cumbersome framework. The second is that the agencies generating the data update and improve the methodologies of data collection and estimation at irregular time periods. This would create a dependence on the agencies to maintain the model and react to their future unknown changes.

As in most classification algorithms we predefine the number of groups we have in the process, in our case we split the rate of growth into three groups: low, mid and high. Assuming there are varying revision rate distributions under different growth regimes, the one-sided Mann-Whitney-Wilcoxon (MWW) test is applied to evaluate the maximum median difference of the revision rates between the low and high process growth, at 95% confidence level. Where the change in growth rate between two consecutive activity periods is evaluated for each revision. We use a line search on the support of each revision in order to locate the largest distance between the low and high

growth regimes. The parameter  $\vartheta \in (0.05, 0.45)$  is used to split the revision distribution into three sections by percentile:

$$\{[p_0 - p_\vartheta], (p_\vartheta - p_{(1-\vartheta)}), (p_{(1-\vartheta)} - p_{100}]|\vartheta\}. \quad (22)$$

The median difference between the revision distribution in the two extreme groupings,  $[p_0 - p_\vartheta]$  and  $(p_{(1-\vartheta)} - p_{100}]$ , and the p-value are evaluated for each revision maturity and level of  $\vartheta$ . After which the means of these measures are calculated across revisions, to create a score for each level of  $\vartheta$ . We define the groups' boundaries as the value of  $\vartheta$  that minimizes the mean median difference across all maturities.

Empirically we find that this non-parametric approach reproduces the same grouping outcomes as the unsupervised fuzzy competitive learning (UFCL) algorithm, derived by Chung & Lee (1994), using the absolute element-wise difference as the dissimilarity metric. Using the MWW test we get an added value of estimating the level of confidence we have in the group differences both overall and for each given maturity.

Using these regime boundaries initial parameter estimates are evaluated using a first order autoregressive model with an exogenous covariate. In equation 3 the exogenous covariate is defined as the revised estimate of the previous activity period. Due to the use of the EM algorithm to infer the state of growth a number of initial parameter estimates are needed to avoid a local solution. We apply a block bootstrap on the autoregressive model to create these sets of initial parameters.

## 5 Case Study: Israel GDP Growth

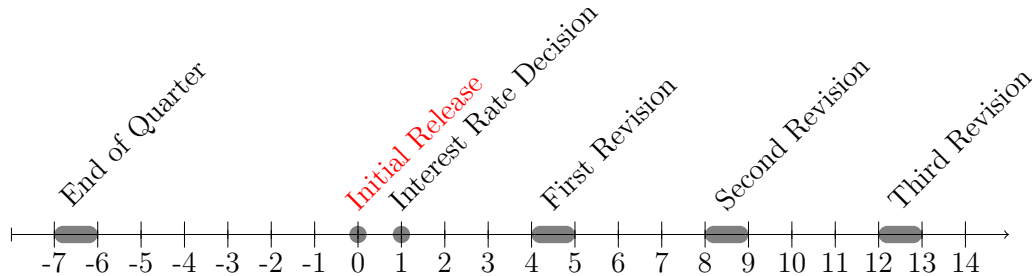
As an illustration of this method, consider the quarterly national accounts data which are published once a month by the Israeli Central Bureau of Statistics (CBS). Each quarterly data report is revised after the initial publication, because of new information and seasonal adjustments. Within the 40 months following the initial publication new information is received by the CBS through surveys and financial reports. Once this information has been taken into account in the revised publication, the full series is once again seasonally adjusted.

Once a month the Bank of Israel (BoI) reassesses its monetary policy stance and a lot of weight is put on the marginal series of the GDP disregarding the divergence between the current GDP publication and the ‘full information’ GDP. This exercise attempts to characterize this divergence on the basis of vintage series of GDP data and construct a prediction interval to the current publication of the seasonally adjusted GDP. The objective of the prediction interval is to broaden the discussion of economic growth from a point estimate constructed from ‘partial information’ found in the initial publications to an estimation of the ‘full information’ GDP, which is found in the revised GDP. Figure 2 shows a timeline centered at the time period of GDP initial publication in relation to the end of quarter activity and the subsequent revisions.

Previous work in the BoI, competitive learning (BOI), has tested the revision convergence rate of the GDP and the possibility of a revision bias with respect to the long term growth trend. The latter test was conditioned on whether the GDP growth being above or below the long run growth trend measured by the HP filter. The findings of this work were that the GDP is updated within 12 periods from initial publication and there was no



Figure 2: Publication Time line in Weeks Centered Around Initial Release



significant bias in the revisions with relation to the long term growth trend.

This case study applies the methods described in the previous sections and constructs a prediction interval for the next revision of GDP growth around the last available estimate of GDP growth. This prediction interval is formulated as a function of the change in GDP growth rate. We find that the revisions are distributed differently depending on the state of the rate of growth. To depict different rates of the growth the revision distributions were examined in three different stages of the economic growth rate cycle: negatively accelerating growth rate, stable growth rate and positively accelerating growth rate. We find evidence that if the rate of growth is negatively accelerating, then the initial publication will be overestimated and subsequent revisions will lower the GDP, and conversely when the rate of growth is positively accelerating the initial publication is underestimated and the subsequent revisions will be higher. Out of sample testing shows that the prediction interval produced at the time of initial release correctly bounds the initial revision of the GDP at a rate of 80% and the future third revision at a rate of 78%.

## 5.1 Characteristics of the GDP Process

The first and second difference of the Israel GDP are depicted in Figure 3. The GDP growth rate and the change in growth rate through boxplots connected at the median value of each quarter in order to visualize the publication volatility. It was found that the revision volatility increases as the absolute value of the GDP rate of growth increases, and that the horizon for volatility convergence is longer with the increase of GDP rate of growth. When testing the convergence rate of the GDP we find that, at a significance of 95%, after 40 revisions following the initial publication of a quarter the volatility of the publications stabilize. This is consistent with findings of other central banks such as the Federal Reserve and the Bank of England, Cunningham et al. (2012), Anderson & Gascon (2009) and Jacobs & Van Norden (2011).

Applying the non-parametric test to differentiate the different revision distribution as a function of the growth process defined in section 4 we find that there is sufficient evidence that validates the existence of three regimes. As seen in Table 1 the median in the first revision of the High Regime is larger than the Low Regime with a P-Value of 5.7%, the fourth revision 6.9%, seventh 0.01% and tenth 7.5% for the median difference. These maturities with at most 10% significant median difference coincide with initial revision and the publication of subsequent activity periods. This result gives validation for the construction of a prediction interval on the current publication GDP while taking into account the growth rate regime. Figure 1 shows for each revision the distribution of each state of growth separated by the  $\vartheta$  that was found to create the greatest distance between the low and high growth regimes.

Figure 3: Distribution of Vintage GDP growth rate and change in growth rate

Sample Period: 2005m9-2014m3

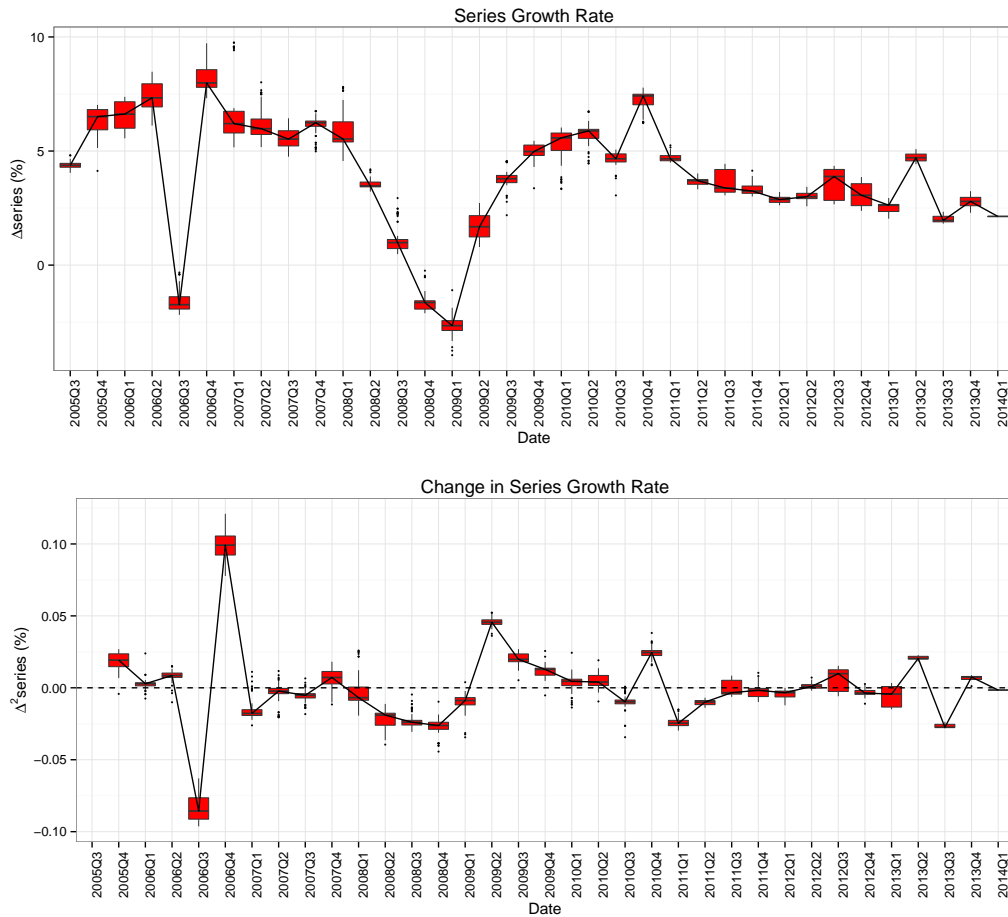


Table 1: Wilcoxon Test for Location Difference Between High Growth and Low Growth Regimes First 12 Revisions, Sample Period 2005m9-2014m3

Test	Upcoming Revision											
	1	2	3	4	5	6	7	8	9	10	11	12
Wilcoxon	0.05	0.70	0.14	0.07	0.75	0.47	0.01	0.71	0.29	0.08	0.63	0.27

## 5.2 HMM estimation

Given a vintage of the published GDP we estimate the probability to be in a given state in regards to the rate of growth. The ARIMAX switching model, was fitted to the data by maximum likelihood

$$y_t^k = a_{0j} + a_{1j}y_{t-1}^k + a_{2j}y_{t-1}^{k+m} + \epsilon_t, \quad \epsilon_t \stackrel{iid}{\sim} N(0, \sigma_{aj}^2), \quad (23)$$

where the states are presumed to follow a three-state Markov chain,  $j = \{\text{Low, Stable, High}\}$ , with transition probabilities  $p_{ij}$ . Maximum likelihood parameters are reported in Table 2. The estimated average rate of growth for each state  $a_{0j}/(1 + a_{1j})$  is -1.46, -0.35, 8.17 respectively. While the covariates of the previous period do not have high significance in terms of Pvalue, the inclusion of this information in the model improves the identification of which state the rate of growth is currently in. The ability to correctly identify the states receives a higher priority over inference in this research since we are using the HMM model to correctly choose which revision distribution is used to create the prediction intervals.

In Table 3 we see that the probability to persist in high growth is low 27%, while the persistence is higher in the low (52%) and stable (63%) states. This gives insight to the probability sustained low growth or contraction over consecutive activity periods, and that after high rate of growth we can expect to rebound in the following period with a low rate of growth.

Figure 4 shows the initial publication of the GDP Rate of Growth Process for each time period. The horizontal lines show the boundaries of the regimes as estimated using the MWW test as described in the previous section. The labels for each time period display the smoothed probability estimates that correspond to the state that has the highest probability, i.e.

Table 2: Maximum Likelihood Estimates of Parameters for Markov-Switching Model of Israel GDP

State	Coefficient	Estimate	Standard Error
Low	$a_0$	-2.39	0.48
Stable	$a_0$	-0.28	0.08
High	$a_0$	2.94	0.54
Low	$a_1$	0.64	0.65
Stable	$a_1$	-0.19	0.13
High	$a_1$	-0.64	0.52
Low	$a_2$	-0.88	0.74
Stable	$a_2$	0.05	0.14
High	$a_2$	-0.18	0.58
Low	$\sigma_a$	2.38	
Stable	$\sigma_a$	0.08	
High	$\sigma_a$	2.44	

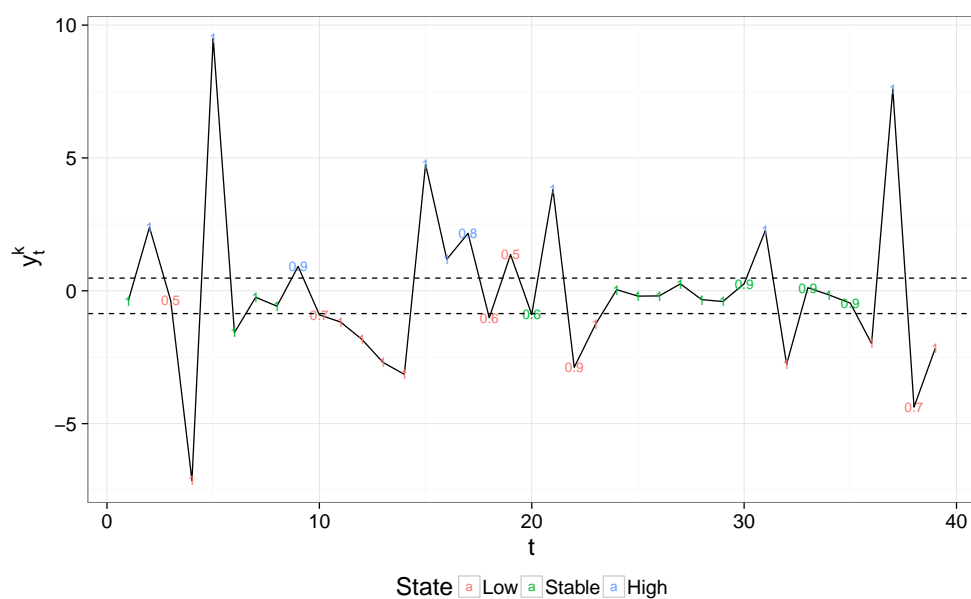
$\max_j \xi_{jt}^0$ , where  $j=\{\text{low,stable,high}\}$ . The model correctly identifies the presumed states of rate of growth in 90% of the periods. We measure the accuracy of the smoothed state probability estimates using a quadratic probability loss function, Brier (1950), defined in Formula 24. The dichotomous variable  $o_{jt}$  denotes that presumed state at time  $t$ , receiving 1 in the presumed state and zero otherwise. This score evaluates to 7%, which indicates a high level of model state prediction accuracy.

$$BS = \frac{1}{T} \sum_{t=1}^T \sum_{j=1}^N (\xi_{jt}^0 - o_{jt})^2 \quad (24)$$

Table 3: State Israel GDP Rate of Growth Probability Transition Matrix

		$S_{t-1}^0$		
		Low	Stable	High
$S_t^0$	Low	0.52	0.13	0.53
	Stable	0.19	0.63	0.20
	High	0.29	0.24	0.28

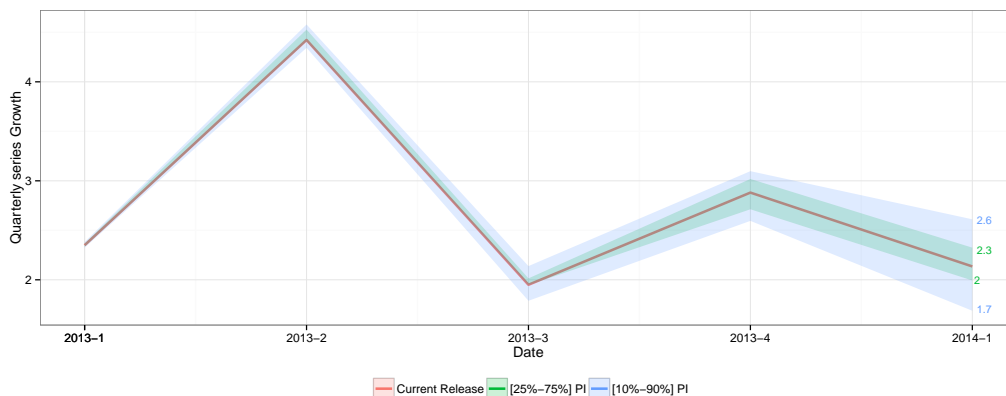
Figure 4: Initial publication ( $k=0$ ) of the GDP Rate of Growth Process ( $y_t^k$ ) and State Probability ( $\max_j \xi_{jt}^k$ , where  $j=\{\text{low,stable,high}\}$ ) for each time period



### 5.3 Updating Current GDP Publication to Next Revision

An example of a prediction intervals can be seen in Figure 5, where the current publication of the GDP includes the initial release of 2014Q1. The initial release of 2014Q1 is estimated to have a prediction intervals of (1.99%,2.32%) and (1.69%,2.61%) with 50% and 80% coverage probability respectively in the upcoming publication. The previous quarters prediction intervals are estimated with respect to their vintage, i.e. 2013Q4 at maturity 4, 2013Q3 maturity 7 and so on. Table 4 presents all the estimates and prediction intervals from 2013Q1-2014Q1 and how many revisions each quarter has had up to this date.

Figure 5: Current GDP Publication and Upcoming Release Revision Range



### 5.4 Out of Sample Testing

An out of sample test was conducted to measure the ability of the algorithm to correctly identify the bounds of an upcoming revision. The algorithm was applied to an expanding window from 2011Q4 to 2014Q1 in which the input

Table 4: Current GDP Publication and Upcoming Release Revision Range

Publication Number	Quarter	Current GDP	Upcoming Revision			
			10%	25%	75%	90%
13	2013Q1	2.35%	2.32%	2.33%	2.37%	2.38%
10	2013Q2	4.42%	4.34%	4.41%	4.53%	4.58%
7	2013Q3	1.95%	1.79%	1.95%	2.01%	2.14%
4	2013Q4	2.88%	2.6%	2.71%	3.02%	3.1%
1	2014Q1	2.13%	1.69%	1.99%	2.32%	2.61%

was the initial publication of the GDP for each quarter. The algorithm was set to create the bounds for the initial revision of the relevant quarter. The measure used for procedure performance is the percent of actual data points that fall within the estimated prediction intervals. In addition a similar measure was calculated for the third future revision, this was done to see if the algorithm could cover up to three months of future GDP publication which are closer to “true” GDP. As can be seen from Table 5 in the first revision estimation the 50% coverage correctly identified the upcoming revision 70% of the time and the 80% coverage correctly identified 80% of the future first revisions. Regarding the third revision the 50% and 80% coverage correctly identified 66% and 78% respectively.

Table 5: Percent of Future GDP Publications in the Prediction Intervals

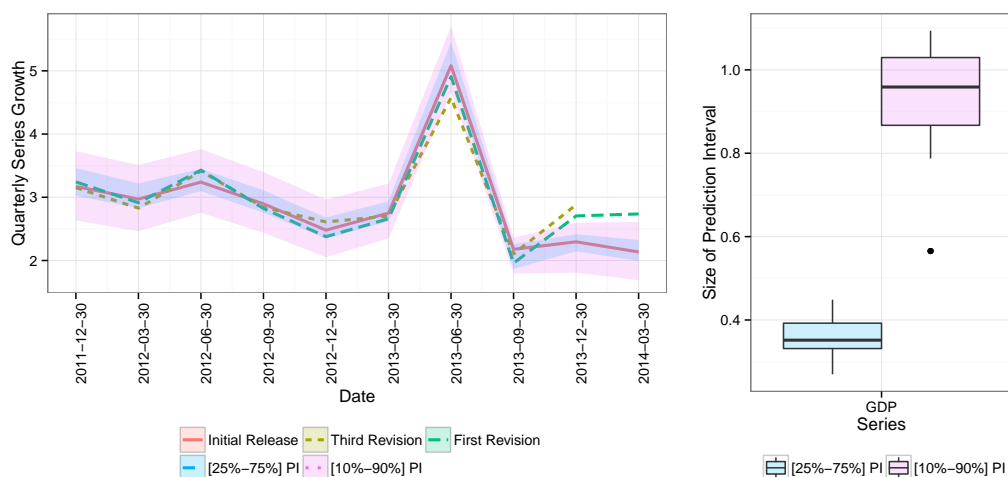
	Prediction Interval	
	[25%-75%]	[10%-90%]
First Revision	70%	80%
Third Revision	66.67%	77.78%



Figure 6 shows the out of sample estimation over the expanding window in the left panel, and the distribution of the prediction interval size in the right panel. We can see that the 50% prediction interval covers the future GDP revisions and that the prediction intervals are tight, i.e. 0.35% for the 50% coverage and 0.95% for the 80% coverage (annualized median growth rate).

Figure 6: Out of Sample One Step Ahead GDP Revision

2011Q4-2014Q1 Seasonally Adjusted Annualized Quarterly Growth



## 6 Conclusion

This paper describes a methodology to generate a prediction intervals for stochastic processes that are continuously revised. This process is characterized as asynchronous, in that random variables in the sequence are revised at each time period, whereas new random variables are added to the sequence at a lower frequency. This creates two levels of uncertainty in the process: the uncertainty between the random variables in the process and uncertainty

---

pertaining to a given random variable which is dependent on the state of the overall process and its maturity since initial estimation.

We proposed in this paper a method to estimate the revision uncertainty with the goal to generate an asymmetrical prediction interval of an upcoming revision of a currently published activity period. These intervals are a function of the state of the process rate of growth. To estimate which state the rate of growth is in at each time point a three-state hidden Markov model was defined.

A case study was conducted on the official publication of the Israel Gross Domestic Product (GDP). We estimated the prediction intervals of the upcoming revision to the current CBS publication of the GDP. This interval characterizes the upper and lower percentiles based on revision distributions from the vintage GDP. It is found that there is a significant difference in the size and sign of revisions dependent on the state of the GDP growth rate. More specifically, when the initial GDP publication is in a low growth quarter, the quarter is overestimated and subsequent revisions lower the GDP, and conversely when the initial GDP publication is in a high growth quarter the quarter is underestimated and the subsequent revisions increase the GDP. This finding was implemented into the calculation of the estimated revision and its prediction interval.

We conclude that this technique constitutes an improvement upon the current point estimates of GDP and GDP growth serving as an input in the assessment of economic activity affecting the monetary policy stance because it provides in addition to this point estimate a prediction interval for the range of fluctuation of the growth rate allowing a more reliable assessment of the strength of economic activity. This approach is novel in comparison current structural models used in leading central banks to estimate confidence

intervals of revisions through the Kalman filter.

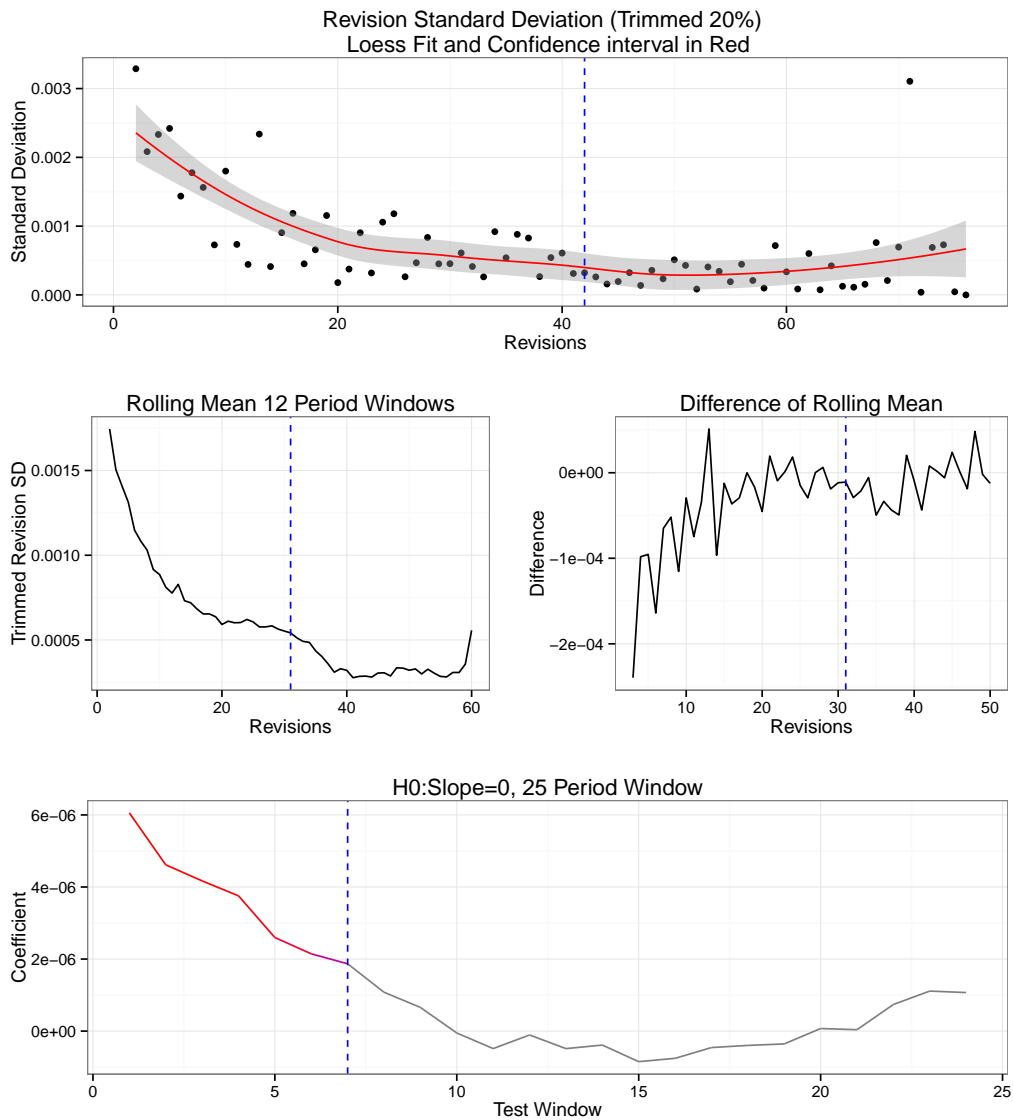
This model framework can be applied to the nowcasting of the GDP. Since the early 2000s the ability to predict the growth rate of an economy within quarter utilizing high frequency data has become common place in major central banks (USA, Canada, Israel, Europe). Usually the nowcast is deployed roughly a month prior to the official publication giving the monetary committees flash estimates of the economic output. These estimates frame the outlook of the committee when deciding on the interest rate updates. When interpreting and using the predictions calculated a very large weight is put on the point estimate. The uncertainty is again implied but not estimated. Since the nowcast GDP consists of one out of sample estimate and is by construction an estimated fit of the actual GDP we can safely assume that the same properties of the actual GDP is found in the nowcast series. Continuing this line we can then apply the prediction interval methodology to the nowcast estimate. This application enhances our estimate and improves the horizon of its effectiveness, where instead of gaining 4 weeks on the official publication we are able to gain estimate 16 weeks using the prediction intervals.

## References

- Revisions to quarterly national accounts data. *Bank of Israel: Recent Economic Developments*.
- Anderson, R. G. & Gascon, C. S. (2009). Estimating us output growth with vintage data in a state-space framework. *Federal Reserve Bank of St. Louis Review*, 91(4), 349–69.

- 
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1), 1–3.
- Chung, F. L. & Lee, T. (1994). Fuzzy competitive learning. *Neural Networks*, 7(3), 539–551.
- Cunningham, A., Eklund, J., Jeffery, C., Kapetanios, G., & Labhard, V. (2012). A state space approach to extracting the signal from uncertain data. *Journal of Business & Economic Statistics*, 30(2), 173–180.
- Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of econometrics*, 45(1), 39–70.
- Jacobs, J. P. & Van Norden, S. (2011). Modeling data revisions: Measurement error and dynamics of true values. *Journal of Econometrics*, 161(2), 101–109.
- Kapetanios, G. & Yates, A. (2004). Estimating time-variation in measurement error from data revisions; an application to forecasting in dynamic models.
- Kim, C.-J. (1994). Dynamic linear models with markov-switching. *Journal of Econometrics*, 60(1-2), 1–22.
- Romer, D. (2012). Dynamic stochastic general equilibrium models of fluctuations. *Advanced Macroeconomics. Fourth ed. New York: McGraw-Hill Irwin*, 312–364.

Figure 7: Revision Decay Analysis



## 7 Appendix

### 7.1 Revision Decay

The revision decay rate was estimated using the seasonally adjusted GDP vintage series 2005m9-2013m8, encompassing the historical monthly series

---

until the change of seasonal adjustment of the GDP known as SNA2013. The trimmed ( $\pm 10\%$ ) standard deviation of each GDP revision was calculated. This can be seen in the first row of Figure 7 along with a loess fit for to visualize a linear decay. When looking at the long term horizon of the revisions one can see that the standard deviation increases once again after 60 revisions (approximately five years after initial release), this is taken as a technical noise due to seasonal adjustments and not new data to the system. The following analysis is done using the first 60 revisions in the vintage GDP data.

To find the horizon length when the standard deviation becomes stable a rolling mean with a 12 period window was calculated to smooth the trimmed standard deviation. Where after the difference between two consecutive rolling windows was calculated,  $\Delta sd(Y_{t*})$ . Finally a rolling regression with a window of 25 periods was applied to test when the slope is not significantly different from zero, we identify the stabilization of the standard deviation when this occurs.

We found that the slope coefficient is not significantly different from zero at window 7 which corresponds to rolling mean windows 7-31. Taking into account the 12 period rolling mean, this gives the time period of 18-42 revisions on the original time scale. The vertical dashed blue line is used as a visual aid to locate these time intervals in Figure 7. One can conclude that up to 42 revisions after the initial publication are needed for the GDP revision to stabilize.

## 4. REGULARIZATION AND CLASSIFICATION OF LINEAR MIXED MODELS VIA THE ELASTIC NET PENALTY WITH APPLICATION TO THE GOOD JUDGMENT PROJECT

**Status:** This chapter was submitted to a peer-review journal and is currently under review. “Supplemental R package CRAN documentation” is given following the main text of the chapter.

---

Regularization and Classification of Linear  
Mixed Models via the Elastic Net Penalty  
with Application to the Good Judgment  
Project\*

Jonathan Sidi<sup>1</sup>, Ya'acov Ritov<sup>1,2</sup> and Lyle Ungar<sup>3</sup>

<sup>1</sup>Department of Statistics, Hebrew University of Jerusalem

<sup>2</sup>Department of Statistics, University of Michigan

<sup>3</sup>Computer and Information Statistics, University of  
Pennsylvania

---

\*This research was supported by a research contract to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.



### **Abstract**

Advances in the field of model selection and prediction via regularization has forged the ability of a variety of disciplines to classify and model large-scale data. Widely used methods which apply penalties in classification are the Least Absolute Shrinkage and Selection Operator (LASSO), the Adaptive LASSO and the Elastic Net. These methods have predominately been used to classify problems of Generalized Linear Models (GLM) in which the dependency of the covariance structure is assumed to be independent. This assumption is not commonly met in practical data and the ability to model such dependencies is integral in fitting the data correctly, such data is modeled using Linear Mixed Models (LMM). Recent research applying LASSO and Adaptive LASSO to LMM's has produced promising initial results of identifying both the random and fixed effects found in data, proving both consistency and an oracle optimality. However, an inherent drawback to those variable selection methods is their performance under high correlation between covariates. To overcome this we introduce the Elastic Net penalty to LMM selection. This penalty has been found to reduce the prediction error in data with high correlation between variables; such a characteristic can be utilized in more complex data designs while optimizing the LMM problem. Findings are tested through simulations and a case study using data accumulated in an longitudinal study where probabilistic forecasts are derived from crowd sentiment. The data structure consists of repeated measures and a large number of fixed and random covariates.

## 1 Introduction

Generalized Linear Mixed models (GLMM) Breslow and Clayton (1993) have been applied in a variety of fields to study data designs with between-subject variation. Such designs include longitudinal, repeated measures and clustered data and have been studied thoroughly in the low dimension setting, e.g., Bates (2010) and Searle et al. (1992). In these settings the linear predictor contains in addition to fixed effects, found in Generalized Linear Models (GLM), latent random effects which capture the unique design pertaining to the data. These random effects usually are assumed to have a centered parametric distribution belonging to the exponential family.

Advances in the field of model selection and prediction via regularization, using different penalty terms, has forged the ability of a variety of disciplines to classify and model large-scale data. Widely used methods which apply penalties in classification are the Least Absolute Shrinkage and Selection Operator (LASSO), Tibshirani (1996), the Adaptive LASSO, Zou (2006b), and the Elastic Net, Zou and Hastie (2005). These methods have predominately been used to classify problems of GLM, Friedman et al. (2010) and Van de Geer (2008), in which the dependency of the covariance structure is assumed to be independent. This assumption in practical data is not commonly met and the ability to model such dependencies is integral in fitting the data correctly, such data is modeled using Linear Mixed Models (LMM) and GLMMs.

Recent research Bondell et al. (2010) a modified Adaptive LASSO (M-ALASSO), smoothly clipped absolute deviation (SCAD) to LMMs and have produced results of identifying both the random and fixed effects found in data, proving both consistency and an oracle optimality. Model selection within the generalized linear mixed models framework has been discussed

in Schelldorfer et al. (2011), Fan and Li (2012), Groll and Tutz (2014), Hui et al. (2016b) and Ibrahim et al. (2011). Schelldorfer et al. (2011) and Groll and Tutz (2014) have a drawback that only fixed effects are selected, while Ibrahim et al. (2011) apply either the SCAD or the ALASSO to each effect. Hui et al. (2016b) allow for greater flexibility for different penalty types on the fixed and random effects. It is noteworthy that Ibrahim et al. (2011) tune each penalty term to a different value through the introduction of the  $IC(q)$  criterion, a characteristic not found in the other methods.

This paper proposes a new penalty called the linear mixed model Elastic Net, LMMEN, which is better suited for regularization in highly correlated data. The LMMEN allows for regularization of both the sparsity ( $\ell_1$  norm) and grouping ( $\ell_2$  norm) for the fixed and random effects separately. We believe that this method will allow to better capture the design of real world data when modeling with LMMs. Through simulations and case study we find that the LMMEN out performs comparative methods in three major areas: highly correlated fixed effects data structures, high dimensionality in the fixed effects, i.e.  $p \gg n$ , selection of random effects when the dimension of the covariance matrix is large.

In the following sections of the paper review the basic structure of the Linear Mixed Effects Model 2, the reparameterization of the LMM to allow for penalization of the random effects 3, define the LMMEN penalty 4, prove asymptotic properties of the penalized model 6, define and analyze simulations comparing the LMMEN to various methods 7, discuss a case study in which the LMMEN is applied to real data 8 and end with discussions 9

## 2 Model

The GLMM is defined as having  $m$  subjects in the sample. For the  $i$ th subject the response variable is denoted as  $y_{ij}$  for the  $j$ th observation, where  $j = 1 \dots n_i$  and let  $N = \sum_{i=1}^m n_i$ . The training data  $\mathbf{X}$  can be defined as two groups of covariates: the fixed effects covariates vector denoted as  $x_{ij}$  with dimensions  $p \times 1$  and the random effects covariates vector denoted as  $z_{ij}$  with dimensions  $q \times 1$ .

$y_{ij}$  are assumed to be conditionally independent given the subject-specific random effects,  $\tilde{\mathbf{b}}_i$ , with a conditional mean  $E[y_{ij}|\tilde{\mathbf{b}}_i] = \mu_{ij}$  and a conditional variance  $\text{var}(y_{ij}|\tilde{\mathbf{b}}_i) = \varrho\omega_{ij}^{-1}\nu(\mu_{ij})$ . Where  $\varrho$  is a positive dispersion parameter,  $\omega_{ij}$  is a pre-specified weight, and  $\nu(\cdot)$  is the variance function. The relationship between  $\mu_{ij}$  to  $\mathbf{X}$  is defined as

$$g(\mu_{ij}) = x'_{ij}\beta + z'_{ij}\tilde{\mathbf{b}}_i,$$

where  $g(\cdot)$  is a strictly increasing link function,  $\beta$  is the fixed effects coefficient vector for  $x$  and  $b_i$  is the subject-specific random effects for  $z$ .  $y_{ij}$  are assumed to be independent and of the form  $y_{ij}|b_i \sim F_y$  and  $\tilde{\mathbf{b}}_i$  is assumed to be of the form  $b_i \sim F_b$ . The distributions  $F_y$  and  $F_b$  are predominately assumed to be normal, i.e.:

$$F_y \sim N(\mu_{ij}, \varrho\omega_{ij}^{-1}\nu(\mu_{ij}))$$

$$F_{\tilde{\mathbf{b}}} \sim N(0, \mathbf{D}(\psi)),$$

where  $\psi$  is a  $q \times 1$  vector of variance components in the covariance matrix of the random effects  $\mathbf{D}$ . Under the identity link function with normal distribution we define the LMM

$$\begin{aligned} \mathbf{y}_i &= x'_{ij}\beta + z'_{ij}\tilde{\mathbf{b}}_i + \boldsymbol{\epsilon}_i, \\ \boldsymbol{\epsilon}_i &\sim N(0, \sigma^2 \mathbf{I}_{n_i}). \end{aligned} \tag{1}$$

McCulloch et al. (2011) state that distribution specification may be affected by basic characteristics of the random effects distribution, such as dependence on a covariate or the cluster sample size. For example, the mean or variance of  $F_{\bar{b}}$  depends on a covariate. When the mean of the random effects distribution depends on a covariate, a fundamental relationship is introduced between the covariate and the distribution, potentially creating a serious bias in estimating the form of the relationship between the covariate and the outcome. Heagerty and Kurland (2001) show that the impact of highly unequal variances can lead to substantial bias. Such bias from distribution specification can cause unintended inference when testing between and within cluster covariates.

Schelldorfer et al. (2011) defined the GLMMLASSO, which solves the log likelihood of the LMM problem via a coordinate gradient descent method based on integral approximation (Laplace approximation) and submit the approximated function to numerical optimization. The advantage of integral approximation methods is to provide an actual objective function for optimization, which enables one to perform likelihood ratio tests among nested models and to compute likelihood-based fit statistics. The disadvantage of these methods is the difficulty of accommodating crossed random effects and multiple subject effects, and the inability to accommodate residual effect covariance structures, or even only residual effect over-dispersion. Moreover, the number of random effects should be small for integral approximation methods to be practically feasible. This disadvantage could potentially inhibit the estimation of random effects in a high dimensional data setting. This characteristic is inherent in all the methods that are built to solve the GLMM problem. The penalty term which is used on the approximated likelihood function is the  $L_1$  penalty. The algorithm proposed penalizes only

the fixed effects in the model, thereby estimating the parameters  $\{\beta, \theta, \rho\}$  and predicting the random effects vector  $b$  using those estimates. The size of the tuning parameter is calculated in two steps: first via the AIC criterion to generate a relevant set of variables and secondly via the BIC criterion to select the final set of active fixed effects which is an unbiased estimator of degrees of freedom in linear models. Hui et al. (2016b) use a regularized penalized quasi-likelihood (rPQL) approach, which also approximates the marginal likelihood function, to simultaneously select fixed and random effects. Groll and Tutz (2014) attempt to solve a similar problem, but with the emphasis on the fixed effects selection. They deploy a gradient ascent algorithm in contrast to the coordinate gradient descent method in Schelldorfer et al. (2011).

Fan and Li (2012) introduce a class of variable selection methods for the fixed effects using a penalized profile likelihood, provided that the random effects vector has a nonsingular covariance matrix. This penalized profile likelihood is equivalent to the penalized quadratic loss function of fixed effects readily found in penalized least squared methods, such as LARS Efron et al. (2004). Random effects are selected under the constraint that the dimension of the fixed effects is smaller than the sample size. They describe an iterative solution for high dimensionality of both the fixed and random effect by which of selecting the fixed effects using the penalized least squares by ignoring all random effects to reduce the number of fixed effects to below sample size. Then in the second step, with the selected fixed effects, they select random effects and finally using the selected random effects refine the fixed effects selections.

Bondell et al. (2010) apply linearization (Taylor expansion) to solve the LMM which is more aptly suited in models with correlated errors, a large

---

number of random effects, crossed random effects, and multiple types of subjects. The disadvantages of this approach include the absence of a true objective function for the overall optimization process and potentially biased estimates. The likelihood function is reparameterized via a modified Cholesky decomposition of the random effects covariance structure Chen and Dunson (2003). This augmentation allows for penalties on both the fixed and random effects. The penalty used in the optimization is the Adaptive Lasso, Zou (2006a), which allows for large amount of shrinkage applied to the zero-coefficients while smaller amounts are used for the non-zero ones which then results in an estimator with improved efficiency and selection properties. The level of the tuning parameter is calculated using the BIC criterion.

The regularization penalty method we propose called the linear mixed model elastic net (LMMEN) will extend the regularization characteristics of Bondell et al. (2010) and Hui et al. (2016b). A shared characteristic between the methods is the simultaneous selection fixed and random effects through penalizing the fixed and random effects separately. Our extension allows for greater flexibility by tuning multiple regularization parameters simultaneously, instead of tuning a single penalty parameter value for both effects. We will apply the Elastic Net penalty on both the fixed and random effects estimates to allow for improved performance when there is a high level of correlation among the fixed and random covariates.

The following table compares the different methods discussed and used in the simulations to compare the proposed method. We focus on the different type of penalties each method uses on the fixed and random effects, the criteria used to tune the regularization parameters, what type of approximations are used on the marginal likelihood function and if there is an R package that accompanies the method.

Model	Research	FE Penalty	RE Penalty	Tuning Criterion	LogLik Approx.	R Package
LMM	This Paper	Elastic Net	Elastic Net	BIC	None	lmmen
	Bondell et al. (2010)	M-ALASSO	M-ALASSO	BIC	None	None
GLMM	Groll and Tutz (2014)	LASSO	None	BIC	Laplace	gmmLasso
	Schelldorfer et al. (2011)	LASSO	None	AIC+BIC	Laplace	lmmlasso
	Hui et al. (2016b)	LASSO	group LASSO	BIC/IC(q)	PQL	rpql
		ALASSO	group ALASSO			
		SCAD	group SCAD			
Ibrahim et al. (2011)	SCAD ALASSO	group SCAD group ALASSO	IC(q)	Laplace	None	
Fan and Li (2012)	SCAD	group SCAD	BIC	Local Linear	None	

Table 1: Summary of methods to regularize linear mixed models and generalized linear mixed models. The comparative studies to this paper use the LASSO, a variant of the Adaptive LASSO (ALASSO, M-ALASSO), or smoothly clipped absolute deviation (SCAD) as the fixed effects (FE) penalties. All but one use the grouped extension of the FE penalty as the random effects (RE) penalty, only Bondell et al. (2010) use an grouping penalty on the RE. This paper uses the Elastic Net penalty to capture correlation characteristics between the variables.

### 3 Reparameterization of the Generalized Linear Mixed Model

This paper will utilize the reparameterization of the LMM model initially defined in Chen and Dunson (2003), and used in Bondell et al. (2010) and Ibrahim et al. (2011). The Cholesky decomposition is readily used to facilitate the estimation of the covariance matrix of the random effects. The main drawback of the Cholesky decomposition when applied to feature selection is that random effects can not be eliminated. This is a result of the fact that the covariance matrix depends on all of these parameters from the decomposition. Reparameterization offers a simple design which regularization penalties can



be easily applied to the fixed and random effects simultaneously, thus giving the ability to remove parameters from the random effects covariance matrix. The covariance matrix of the random effects  $\mathbf{D}$  is factorized as follows:

$$\mathbf{D} = \mathbf{\Lambda}\mathbf{\Gamma}\mathbf{\Gamma}'\mathbf{\Lambda}, \quad (2)$$

where  $\mathbf{\Lambda} = \text{diag}(d_1, \dots, d_q)$  is a  $q \times q$  non-negative diagonal matrix with elements proportional to standard deviations of the random effects, and  $\mathbf{\Gamma}$  is a lower triangular matrix that relates to the correlations among the random effects with the  $(l, m)$  elements denoted  $\gamma_{lm}$ . The elements of  $\mathbf{\Lambda}$  are defined as possibly equal zero, thus enabling a subset of random effects to be selected.  $\mathbf{\Lambda}$  and  $\mathbf{\Gamma}$  are identifiable due to the assumption that:

$$d_l \geq 0, \gamma_{ll} = 1, \text{ and } \gamma_{lr} = 0, \text{ for } l = 1, \dots, q; r = l + 1, \dots, q. \quad (3)$$

Applying the modified decomposition (2) to the LMM model (1) the reparameterized LMM is defined, where the covariance matrix of  $b_i$  is a function of  $\mathbf{\Lambda}, \mathbf{\Gamma}$ :

$$\mathbf{y}_i = \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{Z}_i'\mathbf{\Lambda}\mathbf{\Gamma}\mathbf{b}_i + \boldsymbol{\epsilon}_i. \quad (4)$$

$\mathbf{y}_i$  is assumed to have been centered and predictors standardized, such that  $\mathbf{X}_i'\mathbf{X}_i$  and  $\mathbf{Z}_i'\mathbf{Z}_i$  represent correlation matrices, and  $\mathbf{b}_i = (b_{i1} \dots b_{iq})'$  is a  $q \times 1$  vector of independent  $N(0, \sigma^2 \mathbf{I}_q)$ . The covariance matrix of  $\tilde{\mathbf{b}}_i$  is a function of  $\mathbf{d} = (d_1 \dots d_q)'$ , and  $q(q-1)/2$  free elements of  $\mathbf{\Gamma}$ . Finally defining  $\boldsymbol{\phi} = (\boldsymbol{\beta}', \mathbf{d}', \boldsymbol{\gamma}')'$  as a  $k \times 1$  vector of unknown parameters, where  $k = p + q(q+1)/2$ .

## 4 Simultaneous Variable Selection and Estimation via Regularization Penalties

The Adaptive LASSO has been used as the penalty function on the modified LMM by Bondell et al. (2010) due to its oracle qualities. Although, there are

drawbacks to its use, the primary disadvantage is that candidate covariates correlated to variables chosen in the active set are dropped from the final solution. This characteristic has been found to be a drawback in large scale data with grouped covariates. Moreover, when solving the likelihood of the LMM we can see that the fixed and random effects are dependent.

$$L(\boldsymbol{\phi}|\mathbf{y}, \mathbf{b}) = -\frac{N + mq}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (\|\mathbf{y} - \mathbf{Z}(\mathbf{I}_m \otimes \boldsymbol{\Lambda})(\mathbf{I}_m \otimes \boldsymbol{\Gamma})\mathbf{b} - \mathbf{X}\boldsymbol{\beta}\|^2 + \mathbf{b}'\mathbf{b}), \quad (5)$$

with  $\otimes$  denoting the Kronecker product,  $\mathbf{Z}$  is a block diagonal matrix of  $\mathbf{Z}_i$ ,  $\mathbf{I}_m$  is the identity matrix of dimension  $m$ , and  $\boldsymbol{\phi} = (\boldsymbol{\beta}', \mathbf{d}', \boldsymbol{\gamma}')'$ .

To overcome these issues we apply a variation on the Elastic Net penalty to the reparameterized likelihood function, equation (5), derived in Bondell et al. (2010). The standard Elastic Net penalty denoted as  $P$ , Friedman et al. (2010), is designed to be applied on a fixed effects model where only  $\boldsymbol{\beta}$  is penalized, as seen in (6) below. In this formulation the problem of collinearity is addressed ( $L_2$  penalty) in conjunction with shrinkage of redundant variables ( $L_1$  penalty).

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in R^{p+1}} \left[ \frac{1}{2N} \sum_{i=1}^N (\mathbf{y}_i - \mathbf{X}'_i \boldsymbol{\beta})^2 + P(\boldsymbol{\beta}) \right] \quad (6)$$

$$P(\boldsymbol{\beta}) = \lambda_2 \sum_{j \in P} \beta_j^2 + \lambda_1 \sum_{j \in P} |\beta_j|.$$

We augment (6), while keeping the overall structure and characteristics of the Elastic Net, i.e., the quadratic structure in the  $L_2$  penalty. The reparameterization of the LMM allows the penalty function,  $\tilde{P}(\boldsymbol{\beta}, \mathbf{d})$ , to be dependent on both the fixed and random effects in the model.

In addition, correlated random effects can be included in the final model selection, whereas in the Adaptive LASSO settings this was not possible, thus overcoming the problematic testing of simultaneous random effects, Chen

and Dunson (2003). The objective function of the LMMEN is defined as the following:

$$Q(\phi|\mathbf{y}, \mathbf{b}) = \|\mathbf{y} - \mathbf{Z}\Lambda\Gamma\mathbf{b} - \mathbf{X}\boldsymbol{\beta}\|^2 + \tilde{P}(\boldsymbol{\beta}, \mathbf{d})$$

$$\tilde{P}(\boldsymbol{\beta}, \mathbf{d}) = \lambda_2^f \sum_{i \in P} \beta_i^2 + \lambda_2^r \sum_{j \in Q} d_j^2 + \lambda_1^f \sum_{i \in P} |\beta_i| + \lambda_1^r \sum_{j \in Q} |d_j|. \quad (7)$$

Where  $\tilde{P}$  and  $Q(\phi)$  denote the penalty applied to the likelihood and the penalized log-likelihood. When the final model is not a mixed effects model, but either a fixed effects or random effects model then the original form of  $P(\beta)$  is applied.

## 5 Tuning Parameters Selection

As with all penalized likelihood methods performance depends directly on being able to choose the appropriate value of the tuning parameters. As seen in equation 7 the penalty  $\tilde{P}(\boldsymbol{\beta}, \mathbf{d})$  contains four regularization parameters that need to be tuned. This is a departure from other methods such as Bondell et al. (2010), Groll (2017), Schelldorfer et al. (2011) that tune a single penalty parameter, and Hui et al. (2016a) and Ibrahim et al. (2011) that tune two penalty parameters one for each type of effect. The former papers use the BIC or an iterative method between AIC and BIC as their tuning criterion, and the latter use predominately the IC(q) criterion to tune the two parameters simultaneously. In this paper we chose to use the BIC, it is known that for a typical linear regression model, it is well understood that the traditional best subset selection method with the BIC can identify the true model consistently Shao (1997), Shi and Tsai (2002). The BIC criterion is defined as in Bondell et al. (2010)

$$BIC_{(\lambda_1^f, \lambda_1^r, \lambda_2^f, \lambda_2^r)} = -2L(\hat{\phi}) + \log(N) \times df_{(\lambda_1^f, \lambda_1^r, \lambda_2^f, \lambda_2^r)}. \quad (8)$$

$L(\hat{\phi})$  is obtained from  $L(\phi)$  using the estimate  $\hat{\phi}$  obtained from the value of the set  $(\lambda_1^f, \lambda_1^r, \lambda_2^f, \lambda_2^r)$ . The degrees of freedom are taken as the number of non-zero elements in  $\hat{\phi}$ . The minimal BIC is found through the following method: The penalty ranges of  $(\lambda_1^f, \lambda_1^r, \lambda_2^f, \lambda_2^r)$  are split into discrete sequences, creating a four dimensional grid to search upon. While holding a subset of three penalties constant, the golden section line search, Kiefer (1953), is applied evaluate  $L(\hat{\phi})$  to the target penalty sequence. This is done to each penalty in succession until convergence to the minimal BIC. While this is computationally intensive, the flexibility in selecting specific combinations has higher priority for this research.

## 6 Asymptotics

Assume that the data  $\{(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{y}_i); i = 1 \dots m\}$  is a random sample of  $m$  subjects from a linear mixed-effects model with a probability density function  $f(\mathbf{y}_i | \mathbf{X}_i, \mathbf{Z}_i, \phi)$ . Let  $\mathbf{y}_i$  be an  $n_i \times 1$  response measurements for subject  $i$ ,  $\mathbf{X}_i$  be an  $n_i \times p$  design matrix of explanatory variables, and  $\mathbf{Z}_i$  be an  $n_i \times q$  design matrix of random effects.

Building upon the asymptotic derivation in Bondell et al. (2010), we relax the assumption of  $m > p$  and  $m > q$ , which is a low dimensional problem in both fixed and random effects to  $m < p$  and  $m > q$  which is a high dimensional setting for the fixed effects and a low dimensional for the random effects. In the following theorems we prove that the LMMEN penalized likelihood estimator can identify the true model with probability tending to one, under high dimensional fixed effects conditions.

Let  $\phi = (\boldsymbol{\beta}', \mathbf{d}', \boldsymbol{\gamma}')'$  be a vector of size  $k \times 1$ , where  $\boldsymbol{\beta} \in \mathbb{R}^p$ ,  $\mathbf{d} \in \mathbb{R}^q$  and  $\boldsymbol{\gamma}$  is of the dimension  $\frac{q(q-1)}{2}$ .  $p = m^\alpha$  is the number of fixed effects, and

$q = m^\delta$  the number of the random effects to be estimated. Then number of free elements in the covariance matrix of the random effects,  $\Phi$ , is  $\frac{q(q-1)}{2}$ .

In Bondell et al. (2010) the hyperparameters satisfy  $\alpha < 1$  and  $\delta < 1$  giving a setup of  $m > p$ ,  $m > q$ . The total number of unknown hyperparameters is  $k = p + \frac{q(q+1)}{2} \ll m$ . In this paper we are letting  $\alpha > 1$ ,  $\delta < 1$  giving a framework of  $m < p$ ,  $m > q$ , i.e. a high-dimensional problem. The total number of unknown parameters that are estimated in this framework is  $k = m^\alpha + \frac{m^\delta(m^\delta+1)}{2} \gg m$ .

Let  $L_i(\phi) = \log(f(\mathbf{y}_i | \mathbf{X}_i, \mathbf{Z}_i, \phi))$  denote the contribution of observation  $i$  to the log-likelihood function, given by:

$$L_i(\phi) = -\frac{1}{2} \log |\mathbf{V}_i| - \frac{1}{2} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})' \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}), \quad (9)$$

where  $\mathbf{V}_i = \sigma^2 (\mathbf{Z}_i \boldsymbol{\Lambda} \boldsymbol{\Gamma} \boldsymbol{\Gamma}' \boldsymbol{\Lambda} \mathbf{Z}_i + \mathbf{I}_{n_i})$ . Denoting the true value of  $\phi$  as

$$\boldsymbol{\phi}_0 = (\varphi_{10}, \dots, \varphi_{k0})' = (\boldsymbol{\phi}'_{10}, \boldsymbol{\phi}'_{20})',$$

where  $\boldsymbol{\phi}_{10} = (\boldsymbol{\beta}'_{10}, \mathbf{d}'_{10}, \boldsymbol{\gamma}'_{10})'$  is an  $s \times 1$  vector whose components are non-zero and  $\boldsymbol{\phi}_{20}$  are the  $(k - s)$  remaining components of  $\boldsymbol{\phi}_0$  such that  $\boldsymbol{\phi}_{20} = 0$ . Accordingly, let  $\boldsymbol{\phi} = (\boldsymbol{\phi}'_1, \boldsymbol{\phi}'_2)'$ . To present the theorems the following regularity conditions are imposed:

**C1** The Fisher information matrix  $I(\boldsymbol{\phi}_{10})$  knowing  $\boldsymbol{\phi}_{20} = 0$  is finite and positive definite.

**C2** There exists an open subset  $\Theta$  of  $\mathbb{R}^k$ , containing the true parameter  $\boldsymbol{\phi}_0$  such that  $L_i(\boldsymbol{\phi})$  given in (9) admits all third order derivatives, which are continuous and bounded. There exists a finite mean function  $M_{jlm}(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)$  such that

$$\left| \frac{\partial^3}{\partial \beta \partial \varphi_l \partial \varphi_m} L_i(\boldsymbol{\phi}) \right| < M(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i).$$

We have:

**Theorem 1.** Let  $\phi_0 = (\phi'_{10}, \mathbf{0}')'$ , and the observations follow the LMM model satisfying conditions C1 and C2. If  $w_m \tilde{P}(\boldsymbol{\beta}, \mathbf{d}) \rightarrow 0$ ,  $(\lambda_1^f + \lambda_1^r)\sqrt{s}/mw_m \rightarrow 0$ ,  $(\lambda_2^f + \lambda_2^r)/m \rightarrow 0$ , and  $(\lambda_2^f + \lambda_2^r)s/mw_m \rightarrow 0$ , then there exists a local maximizer  $\hat{\phi} = \begin{pmatrix} \hat{\phi}_1 \\ \mathbf{0} \end{pmatrix}$  of  $Q \left\{ \begin{pmatrix} \hat{\phi}_1 \\ \mathbf{0} \end{pmatrix} \right\}$  such that  $\hat{\phi}_1$  is  $w_m$  consistent for  $\phi_{10}$ .

**Theorem 2.** Let the observations follow the LMM model satisfying conditions C1 and C2. If  $\{\lambda_1^f, \lambda_2^f\} \rightarrow \infty$  and  $\{\lambda_1^r, \lambda_2^r\} \rightarrow \infty$  then with probability tending to 1 for any given  $\phi_1$  satisfying  $\|\phi_1 - \phi_{10}\|_1 \leq Mm^{-1/2}$  and some constant  $M > 0$ ,

$$Q \left\{ \begin{pmatrix} \phi_1 \\ \mathbf{0} \end{pmatrix} \right\} = \max_{\|\phi_2\|_1 \leq Mm^{-1/2}} Q \left\{ \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} \right\}.$$

## 7 Simulations

Simulation testing the model selection performance were carried out on five scenarios. In each scenario 200 data sets were simulated from a multivariate normal density.

$$\mathbf{y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2(\mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i' + \mathbf{I}_{n_i})) \quad (10)$$

The true values of  $(\beta_1, \beta_2) = (1, 1)$ , and the true variance covariance matrix

$$\boldsymbol{\Psi} = \begin{pmatrix} 9 & 4.8 & 0.6 \\ 4.8 & 4 & 1 \\ 0.6 & 1 & 1 \end{pmatrix} \quad (11)$$

The parameterization of the five scenarios are defined in Table 2.

Scenario	Total Obs. N	Subjects m	Obs per subject $n_i$	Fixed Effects (real) p	Random Effects (real) q	Attribute
1	150	30	5	9 (2)	4 (3)	baseline
2	600	60	10	9 (2)	4 (3)	+(N ↑)
3	300	60	5	9 (2)	10 (3)	+(N ↑, q ↑)
4	600	60	10	9 (3)	4 (3)	+(N ↑, $\rho_\beta > 0$ )
5	150	30	5	200 (20)	4 (3)	+(p >> N)

Table 2: Simulation Scenarios in which Scenario 1 is the baseline and each successive scenario builds upon the baseline in a characteristic of interest. Scenario 2 evaluates a larger sample size, Scenario 3 evaluates a larger sample size and more nuisance random effects, Scenario 4 evaluates a larger sample size and correlation between the fixed effect covariates and Scenario 5 evaluates a setting in which there are more fixed effects covariates than total observations.

LMMEN is compared to M-ALASSO Bondell et al. (2010), the R packages that solve: `glmLasso` Groll (2017), `lmlasso` Schelldorfer (2011) and `rPQL` Hui et al. (2016a) which implements the SCAD<sup>1</sup> penalty PQL approximation of the GLMM marginal likelihood function. In all the methods we used the BIC criterion to select the final model in each simulation.

The first three scenarios are taken from Bondell et al. (2010), where the true model under consideration in scenarios 1 and 2 is defined as model (12a) and scenario 3 where  $X = Z$  as model (12b).

$$y_{ij} = b_{i1} + \beta_1 x_{ij1} + \beta_2 x_{ij2} + b_{i2} Z_{ij1} + b_{i3} Z_{ij2} + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, 1) \quad (12a)$$

$$y_{ij} = b_{i1} + (\beta_1 + b_{i2}) x_{ij1} + b_{i3} X_{ij3} + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, 1) \quad (12b)$$

We add two new scenarios to test LMMEN under situations of correlation

<sup>1</sup>The SCAD hyper parameter is set to  $a=3.7$  Fan and Li (2001)

in the fixed effects covariates and when there is a high dimension problem in the fixed effects ( $p \gg n$ ).

Scenario 4 tests the model performance under settings that high correlation between fixed variables exists. We extend scenario 2 by replacing  $X_3$  with a linear combination of  $X_1, X_2$  such that,  $X_3 = wX_1 + (1 - w)X_2 + \epsilon$  where  $\epsilon \sim N(0, \tau)$ . This introduces high correlation in the first three fixed effects, in this setting the LASSO and Adaptive Lasso discard one of these fixed effects thus rendering the model selection inferior. This scenario was found to be beyond the limitations of the glmmlasso and the rPQL packages.

Scenario 5 tests the performance in high dimension settings. The number of fixed effects is increased to 200 and the first 20 are real parameters while the remainder 180 are nuisance, the random effects remain as in the previous scenarios. This scenario can only be run under LMMEN since the initial values are not calculated using the solution of an unpenalized mixed model, as in the other methods.

Tables 3, 4, 5 depict the summary statistics of each parameter estimated within each scenario, where the fixed effects are in Tables 3, 4 and the standard deviations of the random effects are in Table 5. The first three scenarios the real fixed effects are chosen consistently in all the method, where nuisance parameters are chosen with higher regularity. We also notice that the LMMEN and the M-ALASSO coefficient estimates are comparatively underestimated. Scenario 4 we see that LMMEN out performs the other methods by estimating the real parameters including the  $\beta_3$  which is the linear combination of the first two. As expected the other methods discards one of the three parameters due to the use of only  $\ell_1$  penalty. Scenario 5 shows that the LMMEN estimates all real parameters to an weighted average of 0.2 and sets the nuisance to zero on average. In the random effects selection we see



that the LMMEN bias in the estimation of the variance components increases with higher variance levels. The glmLasso, lmmlasso and the rPQL with the SCAD penalty in showed diminished ability to set nuisance parameters to zero.

Table 6 shows the percent of variables correctly selected for the whole model, only the fixed effects and only the random effects for each scenario. This analysis was carried out in two settings, the first summarizes is if the real parameters are a subset of the final model, this is denoted as ‘subset’, and a stricter summary if only the real parameters were chosen in the final model, denoted as ‘oracle’. In the subset analysis the real parameters are selected at high levels in all methods in the first two scenarios. In scenario 4 results seem positive for the lmmlasso, but when cross referencing the estimated values in Table 3 we see that while all variables are included  $\beta_3$  is very close to zero on average in both the lmmlasso and the M-ALASSO, thereby making it’s inclusion less relevant. In the final scenario all the real parameters were in the final model 0% of the simulations, this is due to the large amount of fixed effects that were included, while their coefficient estimates were on average 0.2. The oracle analysis shows that correctly selecting only the real parameters is a much more difficult task and the LMMEN and the M-ALASSO outperformed the other methods in the first three scenarios, in scenario 4 the LMMEN out performed the M-ALASSO, while the lmmen selected the correct parameters, but with near zero coefficient values.

## 8 Case Study

The LMMEN penalty was tested on high dimensional panel data accumulated as part the Good Judgment Project within the Aggregative Contingent

Scenario	Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$
1	glmLasso	1 (0.91,1.06)	1 (0.93,1.06)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
	LMMEN	0.8 (0.48,0.92)	0.8 (0.50,0.93)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
	lmlasso	1 (0.92,1.07)	1 (0.93,1.06)	0 (-0.07,0.07)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
	M-ALASSO	1 (0.88,1.04)	1 (0.91,1.05)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
	rPQL	1 (0.90,1.13)	1 (0.89,1.09)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
	2	glmLasso	1 (0.98,1.03)	1 (0.97,1.03)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
LMMEN		1 (0.93,1.00)	1 (0.92,1.00)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
lmlasso		1 (0.98,1.03)	1 (0.97,1.03)	0 (-0.02,0.03)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
M-ALASSO		1 (0.93,1.00)	1 (0.93,0.99)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
rPQL		1 (0.97,1.03)	1 (0.96,1.03)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
3		LMMEN	0.8 (0.56,0.96)	0 (0,0.01)	0.9 (0.86,1.01)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
	lmlasso	1 (0.82,1.19)	0 (-0.09,0.11)	1 (0.95,1.06)	0 (-0.05,0.06)	0 (-0.05,0.06)	0 (-0.06,0.05)	0 (-0.06,0.05)	0 (-0.07,0.05)	0 (-0.05,0.07)
	M-ALASSO	0.8 (0.75,0.92)	0 (0,0)	0.8 (0.75,0.89)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
	rPQL	1 (0.74,1.17)	0 (0,0)	1 (0.90,1.09)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
4	LMMEN	0.2 (0,0.46)	0.2 (0,0.49)	0.6 (0.35,0.84)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
	lmlasso	1 (0.98,1.03)	1 (0.97,1.03)	0 (-0.01,0.02)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
	M-ALASSO	0.9 (0.40,0.97)	0.9 (0.52,0.99)	0 (0,1.07)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)

Table 3: Scenarios 1-4 summary statistics of fixed effects selection. Each column represents a coefficient and rows are grouped by scenario. In each cell there is the mean value of the estimate and underneath the lower and upper quantile of the estimated coefficient. Not all methods were able to run on scenarios 3,4 due to package constraints relevant to the method.

Estimation (ACE) Program <sup>2</sup>. The aim of this program is *“to dramatically enhance the accuracy, precision, and timeliness of forecasts for a broad range*

<sup>2</sup>Sponsored by the U.S. Intelligence Advanced Research Projects Activity (IARPA).

Scenario	Method											
5	LMMEN	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$	
		0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
		(0,0.34)	(0,0.25)	(0,0.38)	(0,0.27)	(0,0.41)	(0,0.39)	(0,0.33)	(0,0.30)	(0,0.32)	(0,0.36)	(0,0.36)
		$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	$\hat{\beta}_{14}$	$\hat{\beta}_{15}$	$\hat{\beta}_{16}$	$\hat{\beta}_{17}$	$\hat{\beta}_{18}$	$\hat{\beta}_{19}$	$\hat{\beta}_{20}$	
		0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	
		(0,0.41)	(0,0.25)	(0,0.34)	(0,0.4)	(0,0.33)	(0,0.26)	(0,0.34)	(0,0.37)	(0,0.33)	(0,0.39)	
		$\hat{\beta}_{21-200}$										
		0										
		(0,0)										

Table 4: Scenario 5 summary statistics of fixed effects selection. Due to the large dimension of real fixed effects estimated the values are wrapped in a ribbon and the nuisance parameters are grouped together to save space. In each cell there is the mean value of the estimate and underneath the lower and upper quantile of the estimated coefficient. Not all methods were able to run on scenarios 5 due to package constraints relevant to the method.

*of event types, through the development of advanced techniques that elicit, weight, and combine the judgments of many intelligence analysts.”.* The study is characterized as a longitudinal study where probabilistic forecasts are derived from crowd sentiment.

The Good Judgment team recruited approximately 3,000 users in the first year, which varied in ‘expertise’ and randomly assign them to 3 training groups. These training groups are classified as:

- **A: Control** No training
- **B: Probability** Trained to use probabilistic techniques to compare classes and base rates of the occurrence of events, average across expert opinions and assume that long term trends are consistent.
- **C: Scenario** Trained to break down initial assumption to it’s causal drivers, build confidence intervals and worst case scenarios and combine

Scenario	Method	$\hat{d}_1$	$\hat{d}_2$	$\hat{d}_3$	$\hat{d}_4$	$\hat{d}_5$	$\hat{d}_6$	$\hat{d}_7$	$\hat{d}_8$	$\hat{d}_9$	$\hat{d}_{10}$
1	glimLasso	3 (2.70,3.19)	2 (1.78,2.18)	1 (0.86,1.08)	0.3 (0.27,0.37)						
	LMMEN	2.2 (1.98,2.36)	1.6 (1.43,1.74)	0.8 (0.66,0.93)	0 (0,0)						
	hmlasso	3 (2.69,3.15)	2 (1.82,2.21)	1 (0.87,1.09)	0.2 (0.08,0.24)						
	M-ALASSO	2.8 (2.59,3.01)	1.9 (1.67,2.07)	0.9 (0.79,1.02)	0 (0,0)						
	rPQL	2.4 (2.16,2.68)	2 (1.66,2.14)	0.5 (0.25,0.75)	0.2 (0.11,0.36)						
2	glimLasso	3 (2.82,3.17)	2 (1.88,2.14)	1 (0.93,1.06)	0.2 (0.18,0.21)						
	LMMEN	1.8 (1.69,1.87)	1.4 (1.31,1.49)	0.9 (0.79,0.93)	0 (0,0.09)						
	hmlasso	3 (2.81,3.17)	2 (1.88,2.14)	1 (0.93,1.06)	0.1 (0.03,0.09)						
	M-ALASSO	2.1 (2.2,2.1)	1.6 (1.47,1.67)	0.9 (0.81,0.94)	0 (0,0)						
	rPQL	2.9 (2.75,3.08)	2 (1.85,2.13)	0.9 (0.80,0.97)	0.1 (0.09,0.19)						
3	LMMEN	1.9 (1.79,2.12)	1.4 (1.29,1.5)	0.5 (0,0.71)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
	hmlasso	2.9 (2.75,3.19)	2 (1.88,2.14)	1 (0.93,1.09)	0.2 (0.07,0.25)	0.2 (0.09,0.24)	0.2 (0.11,0.28)	0.2 (0.12,0.27)	0.2 (0.12,0.24)	0.2 (0.12,0.28)	0.2 (0.12,0.27)
	M-ALASSO	2.2 (2.06,2.33)	1.6 (1.52,1.74)	0.7 (0.64,0.79)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)
	rPQL	2.2 (1.94,2.45)	1.8 (1.64,2.02)	0.4 (0.25,0.60)	0.2 (0.09,0.3)	0.2 (0.08,0.31)	0.2 (0.09,0.31)	0.2 (0.08,0.29)	0.2 (0.08,0.29)	0.2 (0.09,0.29)	0.2 (0.07,0.27)
4	LMMEN	2.2 (1.98,2.39)	1.6 (1.41,1.76)	0.8 (0.63,0.93)	0 (0,0)						
	hmlasso	3 (2.81,3.17)	2 (1.88,2.14)	1 (0.93,1.06)	0.1 (0.02,0.09)						
	M-ALASSO	2.5 (2.27,2.67)	1.7 (1.55,1.83)	0.8 (0.72,0.98)	0 (0,0)						
5	LMMEN	2.3 (1.99,2.47)	1.7 (1.43,1.99)	0 (0,0.69)	0 (0,0)						

Table 5: Scenarios 1-5 summary statistics of random effects selection. Each column represents a coefficient and rows are grouped by scenario. In each cell there is the mean value of the estimate and underneath the lower and upper quantile of the estimated coefficient. Not all methods were able to run on scenarios 3,4,5 due to package constraints relevant to the method.

assumptions through their causal drivers.

Within each training group there are 4 types of opinion polls in which a user

Scenario	Method	C (subset)	CF (subset)	CR (subset)	C (oracle)	CF (oracle)	CR (oracle)
1	glmmLasso	1.00	1.00	1.00	0.00	0.79	0.00
	LMMEN	0.86	0.94	0.90	0.49	0.49	0.78
	lmmlasso	1.00	1.00	1.00	0.00	0.00	0.07
	M-ALASSO	0.98	1.00	0.98	0.49	0.49	0.98
	rPQL	0.97	1.00	0.97	0.01	0.78	0.01
2	glmmLasso	1.00	1.00	1.00	0.00	0.87	0.00
	LMMEN	1.00	1.00	1.00	0.29	0.29	0.45
	lmmlasso	1.00	1.00	1.00	0.00	0.00	0.07
	M-ALASSO	1.00	1.00	1.00	0.64	0.64	0.93
	rPQL	1.00	1.00	1.00	0.00	0.85	0.00
3	LMMEN	0.65	0.96	0.68	0.19	0.19	0.67
	lmmlasso	0.97	0.97	0.97	0.00	0.00	0.00
	M-ALASSO	0.54	0.54	0.98	0.47	0.47	0.77
	rPQL	0.79	0.80	0.97	0.00	0.62	0.00
4	LMMEN	0.38	0.41	0.88	0.26	0.26	0.86
	lmmlasso	1.00	1.00	1.00	0.07	1.00	0.07
	M-ALASSO	0.07	0.07	0.98	0.04	0.04	0.94
5	LMMEN	0.00	0.00	0.29	0.00	0.00	0.23

Table 6: Scenarios 1-5 percentage of datasets with correctly selected parameters. Two types of selection criteria are summarized. The left hand side (denoted as subset) checks if at least the real coefficients were selected and the right hand side (denoted as oracle) checks if only the real variables were chosen and no nuisance parameters. ‘C’ denotes correct for the overall model, ‘CF’ denotes correct fixed effects and ‘CR’ denotes correct random effects. The cells contain the ratio of iterations that returned the correct values per column head and method. Not all methods were able to run on scenarios 3,4,5 due to package constraints relevant to the method.

can be assigned. These groups are classified as:

- **1: Independent** Requires forecasters to work independently.
- **2: Crowd Beliefs** Forecasters see the distribution of the group’s fore-

casts.

- **3: Prediction Markets**<sup>3</sup> System prices the bet by offering a contract to the participant that will pay a fixed amount if and only if s/he is correct.
- **4: Teamwork** Groups of 20-25 forecasters who explain why they make their forecasts, view the explanations of others, comment on them, coordinate a division of labor and enforce group beliefs.

These opinion polls allow for different levels of interactions between the users in each group. Thus making users randomly assigned to 12 groups. 75 active questions were opened over the first year. These questions had various themes such as the future outcome of: financial turbulence, election results, economic stability and diplomatic security, a full list of the questions can be found in the Appendix B Table 7. Each user could answer an active question at any time until the question was closed and resolved. This design is a natural one for a repeated measures model with random effects, in which the questions are designated as subjects with random intercepts and for each group a random effect is estimated. In addition there are 40 fixed variables that contain demographic, psychological and past performance information, a full list of the variables can be found in the Appendix B Table 8. The data tested was 100 random samples of 50 answers from 20 randomly sampled questions, giving a block structure of 1,000 observations.

A model defined to predict weights to assign to each user per question

---

<sup>3</sup>This group has been omitted due to technical problems that arose during the first year of the competition.

and aggregating user outcome predictions to a group outcome prediction.

$$\begin{aligned} \frac{y_i}{1 - y_i} &= x'_{ij}\beta + z'_{ij}b_i + \epsilon_i, \\ \epsilon_i &\sim N(0, \sigma^2 \mathbf{I}_{n_i}). \end{aligned} \tag{13}$$

where  $y_i$  is the probability of each true outcome,  $\beta$  is the fixed effects coefficient vector for  $x$  and  $b_i$  is the subject-specific random effects for  $z$ , which are the (training::opinion poll) group designations for each forecaster. We define  $\hat{w}_i$  as the predicted weight. The estimated weight is transformed to better separate predictions that are in the middle the  $[0, 1]$ , which indicates indecision and forecasts that are closer to extremes of the range. To achieve this the weights are transformed by exponentiation  $\tilde{w}_i = \exp(\hat{w}_i)$  and truncated to the 20<sup>th</sup> and 80<sup>th</sup> percentile of the estimated vector.

The LMMEN with specifications for the design structure will be compared to the rPQL method with the SCAD penalty on both the fixed and random effects. Both of the methods are suitable to the data structure to select relevant fixed effects and determine which training/opinion poll groups have non-zero variance parameters. Additionally both methods share the regularization characteristic of penalizing the fixed and random effects separately, but the rPQL uses the same penalty value for both, while the LMMEN allows for greater flexibility. Two levels of algorithm performance will be investigated, first is the model selection and second is the accuracy of the aggregated predictions. The statistic which will be used to test performance of the aggregated predictions is the Brier score. In this case study only binary events are taken under consideration and 6 questions are omitted under this constraint. The Brier score equation is defined as

$$BS = \frac{1}{N} \sum_{t=1}^N (f_t - o)^2,$$

in which  $f_t$  is the prediction at time  $t$ ,  $o$  is the question outcome, and  $N$  is the number of prediction instances. First we compare the fixed effect

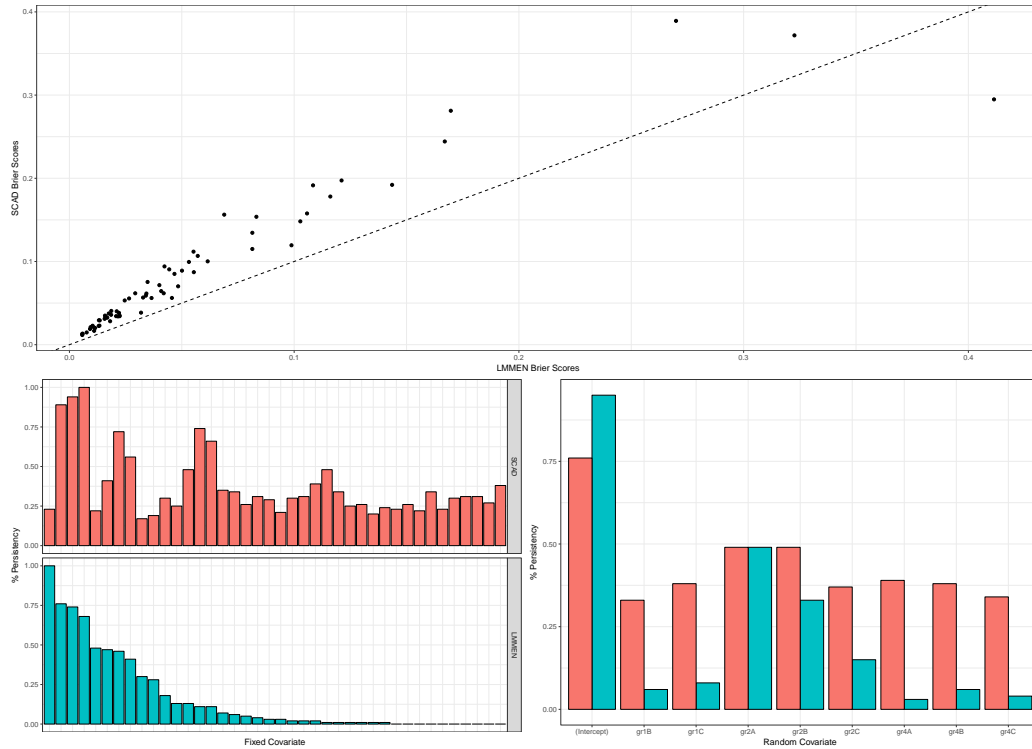


Figure 1: Model performance of LMMEN and rPQL with SCAD penalty tested on 100 random samples of 1000 observations from the Good Judgment study. Panel (a) compares the probabilistic forecast accuracy of the two methods using the Brier Score as the loss function. Panel (b) compares the distribution of fixed effects selection persistency between the two methods. Panel (c) compares the distribution of random effects selection persistency between the two methods.

covariate selection between the two algorithms as seen in panel (b) of Figure 1. It can be seen that the LMMEN produces a higher level of sparsity than the rPQL with SCAD penalty and the variables chosen are persistent in the simulation. The groups of users are assumed to be distributed nor-



mally with a variation parameter. The results of random effects selection can be found in panel (c) of Figure 1. In the LMMEN solution we see that the variance estimate of groups {gr1B: Independent::Probability, gr1C: Independent::Scenario, gr4A: Teamwork::Control, gr4B: Teamwork::Probability, gr4C: Teamwork::Scenario} is equal to zero in a large percent of the simulations, thus concluding that there is no difference between the user responses in those groups, whereas in the rPQL solution there is a some random effects selection but it does not have persistency among any of the groups.

The second level of performance investigated is the prediction accuracy. The estimated non-zero covariates after selection are used to aggregate out of sample user predictions of active questions. The results of the two selection methods can be found in panel (a) of Figure 1. We see that the LMMEN out performs the rPQL in nearly every question, the average Brier scores for LMMEN and rPQL are .055 and .085 respectively.

## 9 Discussion

In the paper we have shown that fixed and random effects in high dimensional linear mixed models can be simultaneously selected. This selection method introduces the ability to select variables under conditions of multicollinearity both in the fixed and random effects. This method, LMMEN, furthers current variable selection of these models with the introduction of a ridge penalty into the optimization.

It was found through simulations that this method correctly selects fixed and random effects under sparse data designs. Simulations were carried out under the Gaussian assumption for both the conditional distribution and the distribution of the random effects. Further simulations are carried out

which relax the assumption of the conditional distribution. We prove that our penalized estimators identifies the true model with probability tending to one, under high dimensional fixed effects conditions. When testing the LMMEN in the case study the variable selection was more apparent both in the fixed and random effects. The LMMEN gave further insight into the characteristics of groups of users, where a subset of them were found not have prediction difference within the groups. Finally, we show that the prediction accuracy of the LMMEN model outperforms the rPQL with a SCAD penalty.

This paper applies the Brier Score (L2 loss) as the loss function to tune the penalty parameters in the case study. One could calibrate the penalty parameters to intraclass correlation (ICC) levels. The ICC is intrinsic to random effects models, and is regularly used for evaluating the level of correlation between different groups as defined by the model. Applying the LMMEN while calibrating to minimize the ICC could be a vital tool for correctly selecting candidate random effects to model the data design and will be assessed in future work.

## 10 Appendix: Proofs

For the penalized log-likelihood in (7), let  $\boldsymbol{\phi} = (\boldsymbol{\phi}'_1, \mathbf{0}')'$  and let

$$L^1(\boldsymbol{\phi}_1) \equiv L \left\{ \begin{pmatrix} \boldsymbol{\phi}_1 \\ \mathbf{0} \end{pmatrix} \right\} \text{ and } Q^1(\boldsymbol{\phi}_1) \equiv Q \left\{ \begin{pmatrix} \boldsymbol{\phi}_1 \\ \mathbf{0} \end{pmatrix} \right\}$$

denote the log-likelihood and the penalized log-likelihood of the first  $s$  components of  $\boldsymbol{\phi}$ .

*Proof Theorem 1.* Consider the penalized log-likelihood  $Q(\boldsymbol{\phi})$  given in (7) in the neighborhood of the true value  $\boldsymbol{\phi}_{10}$ . Let  $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_{10} + \omega_m \mathbf{u}$ , where  $\omega_m = \omega_{m,\alpha} = m^{-\alpha}$ ,  $\forall \alpha > 1$ , and  $\mathbf{u} \neq \mathbf{0}$ . Setting  $\boldsymbol{\phi}_2 = \mathbf{0}$ , we show that for a

small enough  $\epsilon > 0$ , there exists a large constant  $C$  such that for a sufficiently large  $m$ ,

$$P\left(\sup_{\|\mathbf{u}\|=C} Q^1(\boldsymbol{\phi}_{10} + w_m \mathbf{u}) < Q(\boldsymbol{\phi}_{10})\right) \geq 1 - \epsilon.$$

Thus, with probability  $1 - \epsilon$  the maximum is within the ball of radius  $Cw_m$ .

Note that

$$\begin{aligned} mD_m(\mathbf{u}) &\equiv Q^1(\boldsymbol{\phi}_1) - Q^1(\boldsymbol{\phi}_{10}) \\ &= -[L^1(\boldsymbol{\phi}_{10} + w_m \mathbf{u}) - L^1(\boldsymbol{\phi}_{10})] \\ &\quad + \lambda_1^f(\|\boldsymbol{\beta}_0 + w_m \mathbf{u}_\beta\|_1 - \|\boldsymbol{\beta}_0\|_1) + \lambda_1^r(\|\mathbf{d}_0 + w_m \mathbf{u}_d\|_1 - \|\mathbf{d}_0\|_1) \\ &\quad + \lambda_2^f(\|\boldsymbol{\beta}_0 + w_m \mathbf{u}_\beta\|_2^2 - \|\boldsymbol{\beta}_0\|_2^2) + \lambda_2^r(\|\mathbf{d}_0 + w_m \mathbf{u}_d\|_2^2 - \|\mathbf{d}_0\|_2^2), \end{aligned}$$

where we divided  $\mathbf{u}$  to its natural components  $\mathbf{u}_\beta \in R^p$  and  $\mathbf{u}_d \in R^q$ . Using the Taylor series expansion we have

$$\begin{aligned} D_m(\mathbf{u}) &= -w_m(m^{-1}\nabla L(\boldsymbol{\phi}_{10}))' \mathbf{u} - \frac{w_m^2}{2m} \mathbf{u}' [\nabla^2 L(\boldsymbol{\phi}_{10})] \mathbf{u} + R_m \\ &\quad + m^{-1} \lambda_1^f(\|\boldsymbol{\beta}_0 + w_m \mathbf{u}_\beta\|_1 - \|\boldsymbol{\beta}_0\|_1) + m^{-1} \lambda_1^r(\|\mathbf{d}_0 + w_m \mathbf{u}_d\|_1 - \|\mathbf{d}_0\|_1) \\ &\quad + m^{-1} \lambda_2^f(\|\boldsymbol{\beta}_0 + w_m \mathbf{u}_\beta\|_2^2 - \|\boldsymbol{\beta}_0\|_2^2) + m^{-1} \lambda_2^r(\|\mathbf{d}_0 + w_m \mathbf{u}_d\|_2^2 - \|\mathbf{d}_0\|_2^2), \end{aligned}$$

where  $\nabla L(\boldsymbol{\phi}_{10})$ ,  $\nabla^2 L(\boldsymbol{\phi}_{10})$  denote the vector and matrix of the first and second order partial derivatives of  $L(\boldsymbol{\phi}_1)$  at  $\boldsymbol{\phi}_{10}$  respectively.  $\nabla \tilde{P}(\boldsymbol{\beta}, \mathbf{d})$ ,  $\nabla^2 \tilde{P}(\boldsymbol{\beta}, \mathbf{d})$  denote the first and second derivatives of the penalty term at  $(\boldsymbol{\beta}_0, \mathbf{d}_0)$ . The remainder  $R_m$  tends to zero as  $m \rightarrow \infty$  since, by C2,  $|R_m|$  can be bounded by

$$\left(\frac{w_m^3 \|\mathbf{u}\|_2^3}{6m}\right) \sum_{i=1}^m M(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i) = O_P(w_m^3).$$

The  $j$ th partial derivative for each corresponding  $\beta_1, d_1, \gamma_1$  the  $\nabla L(\boldsymbol{\phi}_{10})$  satisfies  $E \left\{ \frac{\partial}{\partial \beta_j} L(\boldsymbol{\phi}_1) \right\} = E \left\{ \frac{\partial}{\partial d_j} L(\boldsymbol{\phi}_1) \right\} = E \left\{ \frac{\partial}{\partial \gamma_j} L(\boldsymbol{\phi}_1) \right\} = 0$  and thus the corresponding empirical means are  $O_p(m^{-1/2})$ .

For  $\nabla^2 L(\boldsymbol{\phi}_{10})$  we have

$$m^{-1} \nabla^2 L(\boldsymbol{\phi}_{10}) \rightarrow_p -I(\boldsymbol{\phi}_{10}),$$

where  $I(\boldsymbol{\phi}_{10})$  is the Fisher information evaluated at  $\boldsymbol{\phi}_{10}$ , which is positive definite by (C1). By choosing a sufficiently large  $C$ , the second term dominates the first term uniformly in  $\|\mathbf{u}\| = C$ .

For the penalty term, if  $w_m \tilde{P}(\boldsymbol{\beta}, \mathbf{d}) \rightarrow 0$  as  $m \rightarrow \infty$  it follows that  $\tilde{P}(\boldsymbol{\beta}, \mathbf{d}) \rightarrow_p 0$ , and thus also dominated by the second term. The absolute value of the penalty component of  $D_m(\mathbf{u})$  is bounded by

$$\begin{aligned} & m^{-1} w_m \lambda_1^f \|\mathbf{u}_\beta\|_1 + m^{-1} w_m \lambda_1^r \|\mathbf{u}_d\|_1 + m^{-1} \lambda_2^f (2w_m \|\boldsymbol{\beta}_0\|_2 \|\mathbf{u}_\beta\|_2 + w_m^2 \|\mathbf{u}_0\|_2^2) \\ & \quad + m^{-1} \lambda_2^r (2w_m \|\mathbf{d}_0\|_2 \|\mathbf{u}_d\|_2 + w_m^2 \|\mathbf{u}_d\|_2^2) \\ & \leq m^{-1} w_m C (\lambda_1^f \sqrt{s} + \lambda_1^r \sqrt{s} + \lambda_2^f (2\|\boldsymbol{\beta}_0\|_2 + w_m C) + \lambda_2^r (2\|\mathbf{d}_0\|_2 + w_m C)). \end{aligned}$$

which is dominated by the second term of  $D_m(\mathbf{u})$ . Therefore, by choosing a sufficiently large  $C$  there exists a local maximum inside  $\{\boldsymbol{\phi}_{10} + w_m \mathbf{u} : \|\mathbf{u}\| < C\}$  with probability  $1 - \epsilon$ , thus there exists a local maximizer  $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\phi}}_1, 0)$  of  $\boldsymbol{\phi}_0 = (\boldsymbol{\phi}_1, 0)$  such that  $\|\hat{\boldsymbol{\phi}}_1 - \boldsymbol{\phi}_{10}\| = O_p(w_m)$ .  $\square$

For the following proof we define  $\boldsymbol{\phi} = (\beta', d', \gamma')$  as a  $k \times 1$  vector of unknown parameters of size  $k = k_\beta + k_d + k_\gamma$ . Let  $\boldsymbol{\phi}_2 = (\beta'_2, d'_2, \gamma'_2)$  be a vector of size  $k_2 = k - s$  corresponding to the true zero parameters, given  $k_2 = k_{\beta_2} + k_{d_2} + k_{\gamma_2}$ . Reminding that we defined earlier that the likelihood and the penalized log likelihood as

$$L(\boldsymbol{\phi}) = L \left\{ \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \end{pmatrix} \right\} \text{ and } Q(\boldsymbol{\phi}) = Q \left\{ \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \end{pmatrix} \right\}.$$

*Proof Theorem 2.* For  $m \rightarrow \infty$  and any  $\boldsymbol{\phi}_1 : \|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_{10}\|_1 \leq Mm^{-1/2}$  and for  $\epsilon_m = Mm^{-1/2}$  and for each  $j = (s+1), \dots, (k_{\beta_2} + k_{d_2})$  we have with probability tending to 1 that

$$\begin{aligned} \frac{\partial}{\partial \varphi_j} Q(\boldsymbol{\phi}) &< 0 \text{ for } 0 < \varphi_j < \epsilon_m \\ \frac{\partial}{\partial \varphi_j} Q(\boldsymbol{\phi}) &> 0 \text{ for } -\epsilon_m < \varphi_j < 0 \end{aligned} \quad (14)$$

The partial derivative of  $Q(\boldsymbol{\phi})$  with respect to  $\varphi_j$  is given by:

$$\frac{\partial}{\partial \varphi_j} Q(\boldsymbol{\phi}) = \frac{\partial}{\partial \varphi_j} L(\boldsymbol{\phi}) - (\lambda_1 \text{sgn}(\varphi_j) + 2\lambda_2 \varphi_j),$$

noting that the penalty is dependent on whether  $\varphi_j$  is  $\boldsymbol{\beta}$  or  $\boldsymbol{d}$ .

One can verify (14) through the Taylor Series expansion of  $\frac{\partial}{\partial \varphi_j} L(\boldsymbol{\phi}) = \frac{\partial}{\partial \varphi_j} L(\boldsymbol{\phi})$  around  $\boldsymbol{\phi}_0$ :

$$\begin{aligned} \frac{\partial}{\partial \varphi_j} Q(\boldsymbol{\phi}) &= \frac{\partial}{\partial \varphi_j} L(\boldsymbol{\phi}_0) - \sum_{l=1}^k \frac{\partial}{\partial \varphi_l} \left( \frac{\partial}{\partial \varphi_j} L(\boldsymbol{\phi}_0) \right) (\varphi_l - \varphi_{l0}) \\ &+ \frac{1}{2} \sum_{i=1}^m \sum_{l=1}^k \sum_{g=1}^k \frac{\partial^2}{\partial \varphi_l \partial \varphi_g} \left( \frac{\partial}{\partial \varphi_j} L_i(\boldsymbol{\phi}_*) \right) (\varphi_l - \varphi_{l0})(\varphi_g - \varphi_{g0}) \\ &- (\lambda_1 \text{sgn}(\varphi_j) + 2\lambda_2 \varphi_j), \end{aligned} \quad (15)$$

where  $\boldsymbol{\phi}_*$  is on the interval connecting  $\boldsymbol{\phi}$  and  $\boldsymbol{\phi}_0$ . Next we define the first order derivatives needed to numerically solve (15):

$$\begin{aligned} L_{\boldsymbol{\beta}} &= \frac{\partial}{\partial \beta_j} L(\boldsymbol{\phi}_0) = \mathbf{X}'_j \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = O_p(m^{-1/2}) \\ L_{\boldsymbol{d}} &= \frac{\partial}{\partial d_j} L(\boldsymbol{\phi}_0) = \frac{1}{2} [\text{Tr}(\mathbf{V}^{-1} \mathbf{S}^j) + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{V}^{-1} \mathbf{S}^j \mathbf{V}^{-1}) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] = O_p(m^{-1/2}), \end{aligned}$$

where  $\mathbf{S}^j = Z(\frac{\partial}{\partial d_j} \mathbf{D}\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\mathbf{D})\mathbf{Z}'$  and  $\text{Tr}(\mathbf{A})$  is the trace operator on a given matrix  $\mathbf{A}$ . We now define the second order derivatives which follow  $\frac{1}{m} \nabla^2 L(\boldsymbol{\phi})|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0} \rightarrow$

$E_{\phi_1=\phi_{10}}[\nabla^2 L(\phi)]$ , where

$$E[\nabla^2 L(\phi)] = E \begin{bmatrix} L_{\beta\beta} & L_{\beta d} & L_{\beta\gamma} \\ L'_{\beta d} & L_{dd} & L_{d\gamma} \\ L'_{\beta\gamma} & L'_{d\gamma} & L_{\gamma\gamma} \end{bmatrix},$$

$$E[L_{\beta\beta}]_j = -\mathbf{X}\mathbf{V}^{-1}\mathbf{X}$$

$$E[L_{\beta d}]_j = -E[\mathbf{X}'_j(\mathbf{V}^{-1}\mathbf{S}^j\mathbf{V}^{-1})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] |_{\phi=\phi_0} = 0$$

$$E[L_{\beta\gamma}]_j = -E[\mathbf{X}'_j(\mathbf{V}^{-1}\mathbf{T}^j\mathbf{V}^{-1})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] |_{\phi=\phi_0} = 0$$

$$E[L_{dd}]_{jl} = -\text{Tr}(\mathbf{V}^{-1}\mathbf{S}^j\mathbf{V}^{-1}\mathbf{S}^l) |_{\{j \geq (s+1), \phi_j=0\}} = 0$$

$$E[L_{\gamma\gamma}]_{jl} = -\text{Tr}(\mathbf{V}^{-1}\mathbf{T}^j\mathbf{V}^{-1}\mathbf{T}^l) |_{\{j \geq (s+1), \phi_j=0\}} = 0$$

$$E[L_{d\gamma}]_{jl} = -\text{Tr}(\mathbf{V}^{-1}\mathbf{S}^j\mathbf{V}^{-1}\mathbf{T}^l) |_{\{j \geq (s+1), \phi_j=0\}} = 0,$$

where  $\mathbf{T}^j = \mathbf{Z}\mathbf{D}(\frac{\partial}{\partial \gamma_j}\boldsymbol{\Gamma}\boldsymbol{\Gamma}')\mathbf{D}\mathbf{Z}'$ .

Using these partial derivatives we solve (15) first for  $\phi_j = \beta_j$  and then for

$$\phi_j = d_j.$$

$$\begin{aligned} & \frac{1}{\sqrt{m}} \left( \frac{\partial}{\partial \beta_j} Q(\phi) \right) \\ = & \frac{1}{\sqrt{m}} \left[ L_\beta - m \left( \sum_{l=1}^{k_\beta} L_{\beta\beta}(\beta_l - \beta_{l0}) + \sum_{l=k_\beta+1}^{k_d} L_{\beta d}(d_l - d_{l0}) + \sum_{l=k_d+1}^{k_\gamma} L_{\beta\gamma}(\gamma_l - \gamma_{l0}) \right) \right. \\ & + \sum_{i=1}^m \sum_{l=1}^{k_\beta} \sum_{g=k_\beta+1}^{k_d} \frac{\partial}{\partial \beta_g} L_{\beta d}(\beta_l - \beta_{l0})(d_g - d_{g0}) \\ & + \sum_{i=1}^m \sum_{l=1}^{k_\beta} \sum_{g=k_d+1}^{k_\gamma} \frac{\partial}{\partial \beta_g} L_{\beta\gamma}(\beta_l - \beta_{l0})(\gamma_g - \gamma_{g0}) \\ & + \sum_{i=1}^m \sum_{l=k_\beta+1}^{k_d} \sum_{g=k_d+1}^{k_\gamma} \frac{\partial}{\partial \gamma_g} L_{\beta d}(d_l - d_{l0})(\gamma_g - \gamma_{g0}) \\ & + \frac{1}{2} \left( \sum_{i=1}^m \sum_{l=k_\beta+1}^{k_d} \sum_{g=k_\beta+1}^{k_d} \frac{\partial}{\partial d_g} L_{\beta d}(d_l - d_{l0})(d_g - d_{g0}) \right. \\ & \left. + \sum_{i=1}^m \sum_{l=k_d+1}^{k_\gamma} \sum_{g=k_d+1}^{k_\gamma} \frac{\partial}{\partial \gamma_g} L_{\beta\gamma}(\gamma_l - \gamma_{l0})(\gamma_g - \gamma_{g0}) \right) - \left( \lambda_1^f \text{sgn}(\beta_j) + 2\lambda_2^f(\beta_j) \right) \Big], \end{aligned}$$

given  $\|\phi - \phi_0\|_1 \leq Mm^{-1/2}$  then we have

$$\frac{1}{\sqrt{m}} \left( \frac{\partial}{\partial \beta_j} Q(\phi) \right) = - \left( \lambda_1^f \text{sgn}(\beta_j) + 2\lambda_2^f(\beta_j) \right) + O_p(1). \quad (16)$$

For  $\beta_{j0} = 0$  and  $\{\lambda_1^f, \lambda_2^f\} \rightarrow \infty$  the sign of the derivative is completely determined by  $\beta_j$ , more specifically:

$$\begin{aligned} & \text{if } M > \beta_j > 0 \quad \text{then } \frac{\partial}{\partial \beta_j} Q(\phi) < 0 \\ & \text{if } -M < \beta_j < 0 \quad \text{then } \frac{\partial}{\partial \beta_j} Q(\phi) > 0 \end{aligned}$$

Similarly,

$$\begin{aligned}
& \frac{1}{\sqrt{m}} \left( \frac{\partial}{\partial d_j} Q(\phi) \right) \\
&= \frac{1}{\sqrt{m}} \left[ L_d - m \left( \sum_{l=1}^{k_\beta} L_{\beta\beta}(\beta_l - \beta_{l0}) + \sum_{l=k_\beta+1}^{k_d} L_{\beta d}(d_l - d_{l0}) + \sum_{l=k_d+1}^{k_\gamma} L_{\beta\gamma}(\gamma_l - \gamma_{l0}) \right) \right. \\
&\quad + \sum_{i=1}^m \sum_{l=1}^{k_\beta} \sum_{g=1}^{k_\beta} \frac{\partial}{\partial \beta_g} L_{d\beta}(\beta_l - \beta_{l0})(\beta_g - \beta_{g0}) \\
&\quad + \sum_{i=1}^m \sum_{l=k_\beta+1}^{k_d} \sum_{g=k_d+1}^{k_\gamma} \frac{\partial}{\partial \gamma_g} L_{dd}(d_l - d_{l0})(\gamma_g - \gamma_{g0}) \\
&\quad + \sum_{i=1}^m \sum_{l=k_\beta+1}^{k_d} \sum_{g=k_d+1}^{k_\gamma} \frac{\partial}{\partial d_g} L_{d\beta}(\beta_l - \beta_{l0})(d_g - d_{g0}) \\
&\quad + \frac{1}{2} \left( \sum_{i=1}^m \sum_{l=k_\beta+1}^{k_d} \sum_{g=k_\beta+1}^{k_d} \frac{\partial}{\partial d_g} L_{dd}(d_l - d_{l0})(d_g - d_{g0}) \right. \\
&\quad \left. + \sum_{i=1}^m \sum_{l=k_d+1}^{k_\gamma} \sum_{g=k_d+1}^{k_\gamma} \frac{\partial}{\partial \gamma_g} L_{d\gamma}(\gamma_l - \gamma_{l0})(\gamma_g - \gamma_{g0}) \right) - (\lambda_1^r \text{sgn}(d_j) + 2\lambda_2^r(d_j)) \left. \right],
\end{aligned}$$

given  $\|\phi - \phi_0\|_1 \leq Mm^{-1/2}$  then we have

$$\frac{1}{\sqrt{m}} \left( \frac{\partial}{\partial d_j} Q(\phi) \right) = -(\lambda_1^r \text{sgn}(d_j) + 2\lambda_2^r(d_j)) + O_p(1).$$

For  $d_{j0} = 0$  and  $(\lambda_1^r, \lambda_2^r) \rightarrow \infty$  the sign of the derivative is completely determined by  $d_j$ , more specifically:

$$\begin{aligned}
& \text{if } M > d_j > 0 \quad \text{then } \frac{\partial}{\partial d_j} Q(\phi) < 0 \\
& \text{if } -M < d_j < 0 \quad \text{then } \frac{\partial}{\partial d_j} Q(\phi) > 0
\end{aligned}$$

□



# 11 Appendix: Case Study Tables

Id	Question Text	Correct Answer	Date Activated	Date Closed	Duration (days)
1	Will the Six-Party talks (among the US, North Korea, South Korea, Russia, China, and Japan) formally resume in 2011?	No	2011-08-31	2012-01-03	125
2	Will Serbia be officially granted EU candidacy by 31 December 2011?	No	2011-08-31	2012-01-03	125
3	Will the United Nations General Assembly recognize a Palestinian state by 30 September 2011?	No	2011-08-30	2011-10-03	34
4	Will Daniel Ortega win another term as President of Nicaragua during the late 2011 elections?	Yes	2011-08-31	2011-11-09	70
5	Will Italy restructure or default on its debt by 31 December 2011?	No	2011-08-31	2012-01-03	125
6	By 1 December 2011, will the World Trade Organization General Council or Ministerial Conference approve the "accession package" for WTO membership for Russia?	Yes	2011-08-31	2011-12-16	107
7	Will the 30 Sept 2011 'last' PPB for Nov 2011 Brent Crude oil futures exceed \$115?	No	2011-09-06	2011-10-03	27
8	Will the Nikkei 225 index finish trading at or above 9,500 on 30 September 2011?	No	2011-09-06	2011-10-03	27
9	Will Italy's Silvio Berlusconi resign, lose re-election/confidence vote, OR otherwise vacate office before 1 October 2011?	No	2011-09-06	2011-10-03	27
10	Will the London Gold Market Fixing price of gold (USD per ounce) exceed \$1550 on 30 September 2011 (10am ET)?	No	2011-09-06	2011-10-03	27
11	Will Israel's ambassador be formally invited to return to Turkey by 30 September 2011?	No	2011-09-06	2011-10-03	27
12	Will PM Donald Tusk's Civic Platform Party win more seats than any other party in the October 2011 Polish parliamentary elections?	Yes	2011-09-06	2011-10-11	35
13	Will Robert Mugabe cease to be President of Zimbabwe by 30 September 2011?	No	2011-09-06	2011-10-03	27
14	Will Mugata al-Sadr formally withdraw support for the current Iraqi government of Nouri al-Maliki by 30 September 2011?	No	2011-09-06	2011-10-03	27
15	Will peace talks between Israel and Palestine formally resume at some point between 3 October 2011 and 1 November 2011?	No	2011-10-03	2011-11-02	30
16	Will the expansion of the European bailout fund be ratified by all 17 Eurozone nations before 1 November 2011?	Yes	2011-10-03	2011-10-17	14
17	Will the South African government grant the Dalai Lama a visa before 7 October 2011?	No	2011-10-03	2011-10-11	8
18	Will former Ukrainian Prime Minister Yulia Tymoshenko be found guilty on any charges in a Ukrainian court before 1 November 2011?	Yes	2011-10-03	2011-10-11	8
19	Will Abdoulaye Wade win re-election as President of Senegal?	No	2011-10-03	2012-03-26	175
20	Will the Freedom and Justice Party win at least 20 percent of the seats in the first People's Assembly (Majlis al-Sha'b) election in post-Mubarak Egypt?	Yes	2012-01-07	2012-01-24	17
21	Will Joseph Kabila remain president of the Democratic Republic of the Congo through 31 January 2012?	Yes	2011-10-03	2012-02-01	121
22	Will Moody's issue a new downgrade of the sovereign debt rating of the Government of Greece between 3 October 2011 and 30 November 2011?	No	2011-10-03	2011-12-01	59
23	Will the UN Security Council pass a measure/resolution concerning Syria in October 2011?	No	2011-10-03	2011-11-01	29
24	Will the U.S. Congress pass a joint resolution of disapproval in October 2011 concerning the proposed \$5+ billion F-16 fleet upgrade deal with Taiwan?	No	2011-10-03	2011-10-26	23
25	Will the Japanese government formally announce the decision to buy at least 40 new jet fighters by 30 November 2011?	No	2011-10-03	2011-12-01	59
26	Will the Tunisian Ennahda party officially announce the formation of an interim coalition government by 15 November 2011?	No	2011-11-07	2011-11-19	12
27	Will Japan officially become a member of the Trans-Pacific Partnership before 1 March 2012?	No	2011-11-07	2012-03-01	115
28	Will the United Nations Security Council pass a new resolution concerning Iran by 1 April 2012?	No	2012-03-21	2012-04-02	12
29	Will Hamad bin Isa al-Khalifa remain King of Bahrain through 31 January 2012?	Yes	2011-11-07	2012-02-01	86
30	Will Bashar al-Assad remain President of Syria through 31 January 2012?	Yes	2011-11-07	2012-02-01	86
31	Will Italy's Silvio Berlusconi resign, lose re-election/confidence vote, OR otherwise vacate office before 1 January 2012?	Yes	2011-11-13	2011-11-15	2
32	Will Lucas Papademos be the next Prime Minister of Greece?	Yes	2011-11-11	2011-11-11	0
33	Will Lucas Papademos resign, lose re-election/confidence vote, or vacate the office of Prime Minister of Greece before 1 March 2012?	No	2011-12-12	2012-03-01	80
34	Will the United Kingdom's Tehran embassy officially reopen by 29 February 2012?	No	2011-12-12	2012-03-01	80
35	Will a trial for Saif al-Islam Gaddafi begin in any venue by 31 March 2012?	No	2011-12-12	2012-04-02	112
36	Will S&P downgrade the AAA long-term credit rating of the European Financial Stability Facility (EFSF) by 30 March 2012?	Yes	2011-12-14	2012-01-17	34
37	Will North Korea successfully detonate a nuclear weapon, either atmospherically, underground, or underwater, between 9 January 2012 and 1 April 2012?	No	2012-01-09	2012-04-02	84
38	By 1 April 2012, will Egypt officially announce its withdrawal from its 1979 peace treaty with Israel?	No	2012-01-09	2012-04-02	84
39	Will Kim Jong-un attend an official, in-person meeting with any G8 head of government before 1 April 2012?	No	2012-01-09	2012-04-02	84
40	Will Christian Wulff resign or vacate the office of President of Germany before 1 April 2012?	Yes	2012-01-09	2012-02-17	39
41	Will the daily Europe Brent Crude FOB spot price per barrel be greater than or equal to \$150 before 3 April 2012?	No	2012-01-09	2012-04-03	85
42	Will the Taliban begin official in-person negotiations with either the US or Afghan government by 1 April 2012?	No	2012-02-22	2012-04-02	40
43	Will Yousaf Raza Gillani resign, lose confidence vote, or vacate the office of Prime Minister of Pakistan before 1 April 2012?	No	2012-01-23	2012-04-02	70
44	Will Yemen's next presidential election commence before 1 April 2012?	Yes	2012-01-23	2012-02-21	29
45	Will Traian Basescu resign, lose referendum vote, or vacate the office of President of Romania before 1 April 2012?	No	2012-01-23	2012-04-02	70
46	Will the UN Security Council pass a new measure/resolution directly concerning Syria between 23 January 2012 and 31 March 2012?	No	2012-01-23	2012-04-02	12
47	Before 1 April 2012, will South Korea officially announce a policy of reducing Iranian oil imports in 2012?	No	2012-01-23	2012-04-02	70
48	Will Israel release Palestinian politician Aziz Duwayk from prison before 1 March 2012?	No	2012-01-23	2012-03-01	38
49	Will Iran and the U.S. commence official nuclear program talks before 1 April 2012?	No	2012-01-30	2012-04-02	63
50	Will Serbia be officially granted EU candidacy before 1 April 2012?	Yes	2012-01-30	2012-03-02	32
51	Will the IMF officially announce before 1 April 2012 that an agreement has been reached to lend Hungary an additional 15+ Billion Euros?	No	2012-01-30	2012-04-02	63
52	Will Libyan government forces regain control of the city of Bani Walid before 6 February 2012?	No	2012-01-30	2012-02-06	9
53	Will a run-off be required in the 2012 Russian presidential election?	No	2012-01-30	2012-03-05	35
54	Will the Iraqi government officially announce before 1 April 2012 that it has dropped all criminal charges against its VP Thaq al-Hashemi?	No	2012-01-30	2012-04-02	63
55	Will Egypt officially announce by 15 February 2012 that it is lifting its travel ban on Americans currently in Egypt?	No	2012-01-30	2012-02-16	17
56	Will a Japanese whaling ship enter Australia's territorial waters between 7 February 2012 and 10 April 2012?	No	2012-02-07	2012-04-11	64
57	Will William Ruto cease to be a candidate for President of Kenya before 10 April 2012?	No	2012-02-07	2012-04-10	63
58	Will Marine LePen cease to be a candidate for President of France before 10 April 2012?	No	2012-02-07	2012-04-10	63
59	Between 21 February 2012 and 1 April 2012, will the UN Security Council announce any reduction of its peacekeeping force in Haiti?	No	2012-02-21	2012-04-02	41
60	Will Mohamed Waheed Hussain Maish resign or otherwise vacate the office of President of Maldives before 10 April 2012?	No	2012-02-21	2012-04-10	49
61	Will Japan commence parliamentary elections before 1 April 2012?	No	2012-02-21	2012-04-02	41
62	Before 13 April 2012, will the Turkish government officially announce that the Turkish ambassador to France has been recalled?	No	2012-02-21	2012-04-13	52
63	Will Standard and Poor's downgrade Japan's Foreign Long Term credit rating at any point between 21 February 2012 and 1 April 2012?	No	2012-02-21	2012-04-02	41
64	Will Myanmar release at least 100 more political prisoners between 21 February 2012 and 1 April 2012?	No	2012-02-21	2012-04-02	41
65	Will a civil war break out in Syria between 21 February 2012 and 1 April 2012?	No	2012-02-21	2012-04-02	41
66	Will Tunisia officially announce an extension of its current state of emergency before 1 April 2012?	Yes	2012-03-05	2012-03-05	29
67	Before 1 April 2012, will Al-Shaabi Gaddafi be extradited to Libya?	No	2012-03-05	2012-04-02	28
68	Before 1 April 2012, will the Sudan and South Sudan governments officially announce an agreement on oil transit fees?	No	2012-03-05	2012-04-02	28
69	Will Yemeni government forces regain control of the towns of Jibar and Zinjibar in the Arabian Peninsula (AQAP) before 1 April 2012?	No	2012-03-05	2012-04-02	28

Table 7: Case study questions asked to participants.

Id	Type	Variable	Label
1	General Knowledge	gk	General knowledge Score
2		gkk	Adjusted General Knowledge Score
3	Current Question	expertise	User given expertise level in question subject matter (scale 1-5)
4		nuAns	Number of new answers user submitted to question
5		timeTo1	Time that passed from activation of question to answer Opinion Poll 1
6		timeTo2	Time that passed from activation of question to answer Opinion Poll 2
7		timeTo4	Time that passed from activation of question to answer Opinion Poll 4
8		timeToSq1	power(timeTo1,2)
9		timeToSq2	power(timeTo2,2)
10		timeToSq4	power(timeTo4,2)
11	Past Performance	nSuc	Number of total correct answers
12		mSuc	Mean number of total correct answers
13		vSuc	Variance of total correct answers
14	User Demographic	Age	Age of user
15		male	gender of user (boolean)
16	User Psychological	baron	Cognitive Reflection Test and Extended Cognitive Reflection Test (by Jon Baron)
17		closure	Need for Closure
18		cons	Political Philosophies
19		fox	Fox-Hedgehog test
20		grit	Grit test
21		needCog	Need for Cognition
22		numer	Berlin Numeracy
23		open	Actively Open-Minded Thinking
24		raven	Number of correct items Raven's Progressive Matrices
25		ravenTime	Truncated time (in seconds) of submit for Raven's item (3 times the user median)
26		ravPerTime	raven/ravenTime
27		logRavTime	log(ravenTime)
28	reflex	Cognitive Reflection Test (CRT)	
30	PolOr1	World politics remains a jungle in which (to quote Thucydides) the strong do what they will and the weak accept what they must	
30	PolOr2	International institutions increasingly constrain the conduct of nation-states	
31	PolOr3	Economic and population growth are stretching nature to its breaking point. Just when humanity seems to be stretching resources to their limits	
32	PolOr4	humans are ingenious at inventing cost-effective technological fixes that permit economic growth to continue.	
33	PolOr5	The rise of China to superpower status will inevitably entail sharp conflicts with the United States.	
34	PolOr6	The rise of radical Islam will be short lived, and pragmatic forces will prevail in contested areas.	
35	PolOr7	I doubt that global climate change modelers know as much about climate trends as they claim.	
36	PolOr8	European monetary integration should be scaled back sharply.	
37	PolOr9	On political and economic issues, I am more liberal than conservative.	
38	PolOr10	Government should routinely intervene in the economy to achieve fairer outcomes.	
39	PolOr11	Free markets function well with minimal government intervention.	
40	PolOr12	I would rather be wrong in an interesting way than right in an uninteresting way.	

Table 8: Case study candidate fixed effects variables.

## References

- D. M. Bates. lme4: Mixed-effects modeling with r. URL <http://lme4.r-forge.r-project.org/book>, 2010.
- H. D. Bondell, A. Krishna, and S. K. Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4): 1069–1077, 2010.
- N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421): 9–25, 1993.
- Z. Chen and D. B. Dunson. Random effects selection in linear mixed models. *Biometrics*, 59(4):762–769, 2003.
- B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
- J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96 (456):1348–1360, 2001.
- Y. Fan and R. Li. Variable selection in linear mixed effects models. *Annals of statistics*, 40(4):2043, 2012.
- J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.
- A. Groll. *glmmLasso: Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation*, 2017. URL <https://CRAN.R-project.org/package=glmmLasso>. R package version 1.5.1.

- 
- A. Groll and G. Tutz. Variable selection for generalized linear mixed models by  $l_1$ -penalized estimation. *Statistics and Computing*, pages 1–18, 2014.
- P. J. Heagerty and B. F. Kurland. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, 88(4):973–985, 2001.
- F. K. Hui, S. Mueller, and A. Welsh. *rpql: Regularized PQL for Joint Selection in GLMMs*, 2016a. URL <https://CRAN.R-project.org/package=rpql>. R package version 0.5.
- F. K. Hui, S. Müller, and A. Welsh. Joint selection in mixed models using regularized pql. *Journal of the American Statistical Association*, (just-accepted), 2016b.
- J. G. Ibrahim, H. Zhu, R. I. Garcia, and R. Guo. Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2):495–503, 2011.
- J. Kiefer. Sequential minimax search for a maximum. *Proceedings of the American mathematical society*, 4(3):502–506, 1953.
- C. E. McCulloch, J. M. Neuhaus, et al. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science*, 26(3):388–402, 2011.
- J. Schelldorfer. *lmmlasso: Linear mixed-effects models with Lasso*, 2011. URL <https://CRAN.R-project.org/package=lmmlasso>. R package version 0.1-2.
- J. Schelldorfer, P. Bühlmann, and S. Van de Geer. Estimation for high-dimensional linear mixed-effects models using  $l_1$ -penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, 2011.

- 
- S. R. Searle, G. Casella, and C. E. McCulloch. *Variance Components*, volume 242. John Wiley and Sons, New York, 1992.
- J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, pages 221–242, 1997.
- P. Shi and C.-L. Tsai. Regression model selectiona residual likelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):237–252, 2002.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- S. A. Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, pages 614–645, 2008.
- H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006a.
- H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006b.
- H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

# Package ‘lmmen’

**Title** Linear Mixed Model Elastic Net

**Version** 1.0

**Date** 2017-08-14

**Description** Fits (Gaussian) linear mixed-effects models  
for high-dimensional data ( $n \ll p$ ) using the linear mixed model elastic-net penalty.

**Depends** R ( $\geq 3.3.2$ ), lmmlasso

**Imports** quadprog, mvtnorm, glmnet, lme4, utils, stats, glmmLasso

**License** GPL-2 | GPL-3

**Encoding** UTF-8

**LazyData** false

**URL** <https://github.com/yonicd/lmmen>

**BugReports** <https://github.com/yonicd/lmmen/issues>

**NeedsCompilation** no

**RoxygenNote** 6.0.1

**Author** Jonathan Sidi [aut, cre]

**Maintainer** Jonathan Sidi <yonis.sidi@mail.huji.ac.il>

## R topics documented:

cv.glmmLasso . . . . .	101
cv.lmmlasso . . . . .	102
golden_section . . . . .	103
golden_section_2d . . . . .	104
init.beta . . . . .	105
initialize_example . . . . .	106
lmmen . . . . .	107
<b>Index</b>	<b>109</b>

---

`cv.glmLasso`      *Cross Validation for glmLasso package*

---

### Description

Cross Validation for glmLasso package as shown in example xxx

### Usage

```
cv.glmLasso(dat, form.fixed = NULL, form.rnd = NULL, lambda = seq(500, 0,
  by = -5), family = stats::gaussian(link = "identity"))
```

### Arguments

<code>dat</code>	data.frame, containing y,X,Z and subject variables
<code>form.fixed</code>	formula, fixed param formula, Default: NULL
<code>form.rnd</code>	list, named list containing random effect formula, Default: NULL
<code>lambda</code>	numeric, vector containing lasso penalty levels, Default: seq(500, 0, by = -5)
<code>family</code>	family, family function that defines the distribution link of the glmm, Default: gaussian(link = "identity")

### Value

list of a fitted glmLasso object and the cv BIC path

### References

A. Groll and G. Tutz. Variable selection for generalized linear mixed models by L1-penalized estimation. *Statistics and Computing*, pages 1–18, 2014.

[cv function is the generalized form of last example glmLasso package demo file](#)

### See Also

[glmLasso](#)

### Examples

```
## Not run: cv.glmLasso(initialize_example(seed=1))
```

---

`cv.lmlasso`      *Cross Validation for lmlasso package*

---

### Description

Cross Validation for lmlasso package as shown in example xxx

### Usage

```
cv.lmlasso(dat, lambda = seq(0, 500, 5), ...)
```

### Arguments

<code>dat</code>	matrix, containing y,X,Z and subject variables
<code>lambda</code>	numeric, path of positive regularization parameter, Default: seq(0, 500, 5)
<code>...</code>	parameters to pass to lmlasso

### Value

lmlasso fit object

### References

J. Schelldorfer, P. Buhlmann, and S. Van de Geer. Estimation for high-dimensional linear mixed-effects models using L1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, 2011.

### See Also

[lmlasso](#)

### Examples

```
## Not run: cv.lmlasso(initialize_example(seed = 1))
```



---

golden\_section      *Golden section grid search on a lmmen penalty*

---

### Description

Solve for local minimum with one dimensional golden section one of the regularization parameters of the lmmen penalty.

### Usage

```
golden_section(dat, init.beta, pen.effect = "FE.L1", opt.lb = 0,
  opt.ub = 1, opt.maxiter = 100, opt.tol = 0.1, opt.tau = (sqrt(5) -
  1)/2)
```

### Arguments

dat	matrix, matrix that includes y (response), X (population covariates), Z (random effects covariates (not incl random intercept))
init.beta	numeric, initial fixed effects estimates
pen.effect	character, which penalty to search on c('FE.L1', 'RE.L1', 'FE.L2', 'RE.L2'), Default: 'FE.L1'
opt.lb	numeric, start of search interval, Default: 0
opt.ub	numeric, end of search interval, Default: 1
opt.maxiter	numeric, maximum iterations to search, Default: 100
opt.tol	numeric, accuracy value, Default: 0.1
opt.tau	numeric, golden proportion coefficient (~0.618) Default: (sqrt(5) - 1)/2

### Value

lmmen list object including lmmen fit object of min BIC solution and summary statistics from the grid search

### Examples

```
## Not run:
dat <- initialize_example(n.i = 5, n = 30, q=4, seed=1)
init <- init.beta(dat, method='glmnet')
golden_section(dat, init, pen.effect = 'FE.L1')

## End(Not run)
```

---

golden\_section\_2d *Golden section two dimensional grid search on L1 lmmen penalties*

---

### Description

Solve for local minimum with two dimensional golden section on L1 lmmen penalties.

### Usage

```
golden_section_2d(dat, init.beta, l2 = c(1, 1), opt.lb = c(0, 0),
  opt.ub = c(1, 1), opt.maxiter = 100, opt.tol = 0.1, opt.tau = (sqrt(5)
  - 1)/2)
```

### Arguments

dat	matrix, matrix that includes y (response), X (population covariates), Z (random effects covariates (not incl random intercept))
init.beta	numeric, initial fixed effects estimates
l2	numeric, L2 penalty levels Default: c(1, 1)
opt.lb	numeric, start of interval for L1 fixed and L1 random effects, Default: c(0, 0)
opt.ub	numeric, end of interval for L1 fixed and L1 random effects Default: c(1, 1)
opt.maxiter	numeric, maximum iterations to search, Default: 100
opt.tol	numeric, accuracy value, Default: 0.1
opt.tau	numeric, golden proportion coefficient (~0.618) Default: (sqrt(5) - 1)/2

### Value

lmmen list object including lmmen fit object of min BIC solution and summary statistics from the grid search

### Examples

```
## Not run:
dat <- initialize_example(n.i = 5, n = 30, q=4, seed=1)
init <- init.beta(dat, method='glmnet')
golden_section_2d(dat, init)
## End(Not run)
```

---

<code>init.beta</code>	<i>Evaluate fixed effects initial values for lmmen</i>
------------------------	--

---

### Description

Evaluate fixed effects initial values for lmmen via `cv.glmnet` or `lme4`.

### Usage

```
init.beta(dat, method = c("glmnet", "lme4"))
```

### Arguments

<code>dat</code>	data.frame, data to solve initial values
<code>method</code>	character, method to use, <code>c('glmnet','lme4')</code>

### Details

`cv.glmnet` is set to ridge regression.

### Value

numeric

### See Also

[cv.glmnet](#)

### Examples

```
dat <- initialize_example(n.i = 5, n = 30, q=4, seed=1)
init.beta(dat, method='glmnet')
init.beta(dat, method='lme4')
```

---

```
initialize_example Initialize Scenario
```

---

### Description

Create a scenario to run the evaluation functions.

### Usage

```
initialize_example(n.i = 5, n = 30, q = 4, total.beta = 9,
  true.beta = c(1, 1, 1), seed = NULL)
```

### Arguments

<code>n.i</code>	integer, Observations per subject, Default: 5
<code>n</code>	integer, Number of subjects, Default: 30
<code>q</code>	integer, Number of random effects, Default: 4
<code>total.beta</code>	integer, Number of simulated fixed effects, Default: 9
<code>true.beta</code>	numeric, True of fixed effects indicies, Default: c(1,1,1)
<code>seed</code>	integer, set a seed for reproducibility, Default: NULL

### Value

$(n.i*n) \times (1+total.beta+q)$  matrix containing where the subjects index are the matrix row-names

	<b>Description</b>	<b>Parameter</b>	<b>Dimension</b>
	Response	<code>y</code>	$(n.i*n) \times 1$
	Fixed	<code>X</code>	$(n.i*n) \times total.beta$
	Random	<code>Z</code>	$(n.i*n) \times q$

### See Also

[rmvnorm](#)

### Examples

```
initialize_example(n.i = 5, n = 30, q=4, seed=1)
initialize_example(n.i = 10, n = 60, q=4, seed=1)
initialize_example(n.i = 5, n = 60, q=10, seed=1)
```

lmmen

*linear mixed model Elastic Net***Description**

Regularize a linear mixed model with the linear mixed model Elastic Net penalty.

**Usage**

```
lmmen(data, init.beta, frac, eps = 10^(-4), verbose = FALSE)
```

**Arguments**

data	matrix, data
init.beta	numeric, initial values for fixed effects coefficients
frac	numeric, penalty levels for fixed and random effects expressed in ratios. c(L1.fixed,L2.fixed,L1.random,L2.random)
eps	numeric, tolerance level to pass to solve.QP, Default: 10 <sup>(-4)</sup>
verbose	boolean, show output during optimization Default: FALSE

**Details**

$$y_i = x_{ij}^t \beta + z_{ij}^t b_i + \epsilon_i,$$

$$\epsilon_i \sim N(0, \sigma^2 I_{n_i})$$

The lmmen function solves for the following problem.

$$Q(\phi|y, b) = \|y - Z\Lambda\Gamma b - X\beta\|^2 + \tilde{P}(\beta, d)$$

$$\tilde{P}(\beta, d) =$$

$$\lambda_2^f \sum_{i \in P} \beta_i^2 + \lambda_2^r \sum_{j \in Q} d_j^2 +$$

$$\lambda_1^f \sum_{i \in P} |\beta_i| + \lambda_1^r \sum_{j \in Q} |d_j|$$

Where  $\tilde{P}$  and  $Q(\phi)$  denote the penalty applied to the likelihood and the penalized log-likelihood.

When the final model is not a mixed effects model, but either a fixed effects or random effects model then the original form of the Elastic Net penalty is applied.

**Value**

lmmen fit object including

- fixed: estimated fixed effects coefficients
- stddev: estimated random effects covariance matrix standard deviations
- sigma.2: standard error of the model residual effect
- lambda: estimated lower triangle of  $\Lambda$  (correlation of random effects)
- Mean.est: model prediction  $X^t\beta$
- loglike: log likelihood
- df: degrees of freedom
- BIC: Minimum BIC penalty value
- frac: ratio placed on the penalties corresponding to BIC
- Gamma.Mat.RE: estimated  $\Gamma$
- Cov.Mat.RE: estimated random effect covariance matrix
- Corr.Mat.RE: estimate random effects correlation matrix
- solveQP: output of the call to solveQP corresponding to min BIC

**References**

[Prepublished version of the lmmen paper.](#)

**See Also**

[solve.QP](#)

**Examples**

```
dat <- initialize_example(n.i = 5, n = 30, q=4, seed=1)
init <- init.beta(dat, method='glmnet')
lmmen(data=dat, init.beta=init, frac=c(0.8, 1, 1, 1))
```

# Index

`cv.glmLasso`, [101](#)  
`cv.glmnet`, [105](#)  
`cv.lmlasso`, [102](#)

`glmLasso`, [101](#)  
`golden_section`, [103](#)  
`golden_section_2d`, [104](#)

`init.beta`, [105](#)  
`initialize_example`, [106](#)

`lmmen`, [107](#)  
`lmlasso`, [102](#)

`rmvnorm`, [106](#)

`solve.QP`, [108](#)

## 5. DISCUSSION



The main goal of this thesis was to examine the dual role that statistical high dimensional models have of the process on time sensitive informed policy decision making. We investigated how policy scenarios create the need for novel methodology, while also finding that new methodology brought to light more focused policy questions. Issues that were presented were: dimension reduction in a central bank setting to derive the benchmark interest rates under partial information (chapter 2), the derivation of prediction intervals for a continuously revised data stream via a hidden Markov model (chapter 3) and the regularization of linear mixed effects model through a new penalty, the linear mixed effects elastic net. While the design in each chapter is different they all focus on the same scope - enhancement of informed policy decision making in a time sensitive high dimensional setting, through novel statistical methodology.

Chapter 2 concerned dimension reduction, with the goal of prediction, of the current state of a system based on partial data. In this chapter we survey two competing dimension reduction methodologies for the purpose of prediction: unconditional on the variable of interest and conditional on the variable of interest. The chapter revolved around policy decision making in central banks. The core decisions are dependent on real-time data analysis as it is published. The ability to produce precise estimates based on partial information, through canonical models, has evolved with the methodological progress of model selection techniques.

This chapter compares model selection techniques applied in leading central banks, which are predominately based on dynamic factor analysis, with regularization of linear models such as the LASSO Tibshirani (1996) and the Elastic Net Zou & Hastie (2005). The application of nowcasting with the Elastic Net of the Israel GDP yielded more precise and stable results,

compared to the other methods surveyed. Moreover, the dynamic nature of the model allows it to adapt to shocks in the economy producing a more robust model. A distinguishing feature of the Elastic Net is the ability to isolate influential variables which contribute to the real-time assessment. This refinement of the results separates this method from current ones used in nowcasting and allows the model to be a more comprehensive tool in economic policy decisions. We show that there is no significant difference between the forecast performance between the initial official publications and the proposed elastic net nowcast estimates. This outcome has important policy ramifications since the nowcast precedes the initial release publications by 4 weeks time, thereby giving the banking committee valuable information that they are currently missing at time of interest rate adjustments. This added value highlights the contribution of advanced data mining techniques in a policy driven economic setting.

Chapter 3 introduces methodology to generate a prediction intervals for stochastic processes that are continuously revised. This process is characterized as asynchronous, in that random variables in the sequence are revised at each time period, whereas new random variables are added to the sequence at a lower frequency. This creates two levels of uncertainty in the process: the uncertainty between the random variables in the process and uncertainty pertaining to a given random variable which is dependent on the state of the overall process and its maturity since initial estimation.

We propose in this chapter a method to estimate the revision uncertainty with the goal to generate an asymmetrical prediction interval of an upcoming revision of a currently published activity period. These intervals are a function of the state of the process rate of growth. To estimate which state the rate of growth is in at each time point a three-state hidden Markov model is

defined.

A case study is conducted on the official publication of the Israel GDP. We estimate the prediction intervals of the upcoming revision to the current CBS publication of the GDP. This interval characterizes the upper and lower percentiles based on revision distributions from the vintage GDP. It is found that there is a significant difference in the size and sign of revisions dependent on the state of the GDP growth rate. More specifically, when the initial GDP publication is in a low growth quarter, the quarter is overestimated and subsequent revisions lower the GDP, and conversely when the initial GDP publication is in a high growth quarter the quarter is underestimated and the subsequent revisions increase the GDP. This finding was implemented into the calculation of the estimated revision and its prediction interval.

We conclude that this technique constitutes an improvement upon the current point estimates of GDP and GDP growth serving as an input in the assessment of economic activity affecting the monetary policy stance because it provides in addition to this point estimate a prediction interval for the range of fluctuation of the growth rate allowing a more reliable assessment of the strength of economic activity. This approach is novel in comparison current structural models used in leading central banks to estimate confidence intervals of revisions through the Kalman filter, such as Cunningham et al. (2012), Anderson & Gascon (2009) and Jacobs & Van Norden (2011).

Combining the added value found in these two chapters we can formalize a method to derive higher levels of informed policy decision making in a real time setting. Since the nowcast GDP consists of one out of sample estimate and is by construction an estimated fit of the actual GDP we can safely assume that the same properties of the actual GDP is found in the nowcast series. Continuing this line the prediction interval methodology is applied to

the nowcast estimate. This application enhances our estimate and improves the horizon of its effectiveness, where instead of gaining 4 weeks on the official publication, by applying only the methodology in chapter 2, we are able to gain 16 weeks using both the nowcast estimates and the prediction intervals. Together the methodologies proposed in chapters 2 and 3 give policy decision makers both an early signal and the context to understand the to what certainty the estimate will remain within given bounds, given the current state of growth of the economy.

Chapter 4 proposes a new penalty to simultaneously select fixed and random effects in high dimensional linear mixed models. This selection method introduces the ability to select variables under conditions of multicollinearity both within the fixed and random effects. The linear mixed effects elastic net penalty, LMMEN. It expands upon the current variable selection of these models, such as Schelldorfer et al. (2011), Fan & Li (2012), Groll & Tutz (2014), Hui et al. (2016) and Ibrahim et al. (2011), with the introduction of a ridge penalty into the optimization.

It was found through simulations that this method correctly selects fixed and random effects under sparse data designs. Simulations were carried out under the Gaussian assumption for both the conditional distribution and the distribution of the random effects. Further simulations are carried out which relax the assumption of the conditional distribution. When testing the LMMEN in the case study the variable selection was more apparent both in the fixed and random effects. The LMMEN gave further insight into the characteristics of groups of users, where a subset of them were found not have prediction difference within the groups. Finally, we show that the prediction accuracy of the LMMEN model outperforms the rPQL with a SCAD penalty, Hui et al. (2016).

The LMMEN penalty was tested on high dimensional panel data accumulated as part the Good Judgment Project within the Aggregative Contingent Estimation (ACE) Program <sup>1</sup>. The aim of this program is “*to dramatically enhance the accuracy, precision, and timeliness of forecasts for a broad range of event types, through the development of advanced techniques that elicit, weight, and combine the judgments of many intelligence analysts.*”. The study is characterized as a longitudinal study where probabilistic forecasts are derived from crowd sentiment.

Chapter 4 contains an extensive supplemental appendix detailing the *LMMEN* R package, Sidi (2017), that solves the linear mixed optimization problem with the linear mixed model elastic net penalty. We go into greater detail regarding the different types of methods used to solve the optimization problems and discuss the implementation of cross validations used in the simulations for both the LMMEN penalty and the other comparative methods. The *LMMEN* package has been released on the CRAN repository.

In conclusion the synthesis of policy decision making and high dimensional methodology is an evolving relationship that each discipline is pushing the other for novel approaches to solve the realities faced today. This thesis focuses on how statistical methodology can help drive policy decision making in the era of high capacity storage capabilities.

## References

Anderson, R. G. & Gascon, C. S. (2009). Estimating us output growth with vintage data in a state-space framework. *Federal Reserve Bank of St. Louis Review*, 91(4), 349–69.

---

<sup>1</sup>Sponsored by the U.S. Intelligence Advanced Research Projects Activity (IARPA).

- Cunningham, A., Eklund, J., Jeffery, C., Kapetanios, G., & Labhard, V. (2012). A state space approach to extracting the signal from uncertain data. *Journal of Business & Economic Statistics*, *30*(2), 173–180.
- Fan, Y. & Li, R. (2012). Variable selection in linear mixed effects models. *Annals of statistics*, *40*(4), 2043.
- Groll, A. & Tutz, G. (2014). Variable selection for generalized linear mixed models by l1-penalized estimation. *Statistics and Computing*, 1–18.
- Hui, F. K., Müller, S., & Welsh, A. (2016). Joint selection in mixed models using regularized pql. *Journal of the American Statistical Association*, (just-accepted).
- Ibrahim, J. G., Zhu, H., Garcia, R. I., & Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, *67*(2), 495–503.
- Jacobs, J. P. & Van Norden, S. (2011). Modeling data revisions: Measurement error and dynamics of true values. *Journal of Econometrics*, *161*(2), 101–109.
- Schelldorfer, J., Bühlmann, P., & Van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics*, *38*(2), 197–214.
- Sidi, J. (2017). *lmmen: Linear Mixed Model Elastic Net*. R package version 1.0.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

## ACKNOWLEDGMENTS

I would like to gratefully thank my advisor, Ya'acov Ritov, for mentoring and teaching me along the way of my thesis and guiding me while I navigate through as a student. I especially appreciate his willingness to help and to advise me on any matter, be it statistics, politics, sports or just life.

I would like to thank my collaborators listed in Chapters 2 and 4. Gil Dafnai was an amazing research partner that I had a privilege to be paired with at the Bank of Israel under the guidance of Amit Friedman, who always pushed us to be more than regular research students.

I would like to thank Alon Binyamini for convincing me to come back as a researcher to the Bank of Israel and being a good friend and great ambassador for high level econometrics and scuba diving. I am very grateful to Tanya Suhoy who was my statistics mentor throughout my stays at the Bank of Israel, she was my inspiration for problem solving, willingness to learn new methodology and not fearing to be steadfast in her opinions.

I would also like to thank various anonymous referees that reviewed earlier versions of Chapters 3 and 4 for their helpful comments.

I would like to thank the dissertation reviewers for very detailed and useful comments. I believe that the suggestions made, considerably improved the paper.

I would like to thank Ramon Garcia and Jayson Wilbur for being my non-anonymous referees and reviewing the chapters in depth and giving advice on how to substantially improve them.

I would like to thank my fellow students Daniel Nevo and Liron Ravner for being willing sounding boards for the development of the chapters. Additionally, I would like to thank the administrative staff and the faculty members at the statistics department for providing my statistics education, and for keeping their doors open for questions and discussions.

Finally, I would like to thank my family and my parents for their endless support and encouragement. If not for my wife and my best friend Yulia this thesis would have not been written, and to my kids Michal and Dan who fill every day with their love, curiosity and experimentations (in the house).





אתגר שמוצג ומוצע כאן בצורה של הפחתת "עלות" הנדרשת ממעקב וסיקור ישיר וממושך בקבוצות גדולות של אוכלוסייה ניבדקת. היכולת למצע את הידע של אוכלוסייה מגוונת כזו עם מאפיינים שונים ולאגור אותה לתחזית אחת, היתה הנושא של מקרה-מבחן שבו בדקנו את ביצועי ה-LMMEN. ניסוי מבוקר זה צבר נתונים אורכיים (longitudinal), שבהם תחזיות הסתברותיות נגזרות מהסנטימנט של ההמון, במטרה לענות בצורה הטובה ביותר על שאלות כלכליות וגיאו-פוליטיות שונות ובעלות ענין ציבורי.

בהבט כללי, כל הפרקים בעבודה זו מתמקדים בהיבטים שונים של האופן שבו האתגרים הסטטיסטיים הכרוכים בנתונים מורכבים רב-מימדים, מיועדים לסייע לעצב את הכיוון החדש של קבלת החלטות על ידי מעצבי החלטות. כל הפרקים הוגשו לפרסום בכתבי עת, ונסקרים ונבדקים בימים אלה. פרק 2 פורסם כנייר לדיון בסדרה של בבנק ישראל, כאשר גם השיטה המחקרית (מתודולוגיה) המוצגת בו וגם זו המוצגת בפרק 3 כבר מיושמות כיום בבנק ישראל, במסגרת ההערכה החודשית של מצב המשק הישראלי לעדכון ריבית המשק.

נוספת של אי ודאות המחייבות כל הזמן הערכה מחדש. יש כל הזמן להעריך ולקחת בחשבון בכל נקודת זמן ובכל מצב כלשהו, כיצד נקודת הנתונים הנוכחית ובכך גם הסדרה כולה להשתנות בתקופות אחרות הבאות עם הזמן. להשתנות כזו עשויות להיות השלכות שעלולות להביא לשינוי בקביעת מדיניות. אפשר לתאר תרחיש מסוים שבו מוצגת מציאות המתאימה למצב ההווה וגורמת לתגובה בצעדים שיהיו אכן נכונים לאותה נקודת זמן, בעוד שבמבט לאחור הנתונים המייצגים את אותן נקודות זמן בסדרה יעודכנו לאחר מכן וייצגו מציאות שונה, ההופכת את החלטת המדיניות שנלקחה ללא מדויקת, עם השלכות אפשריות שליליות כתוצאה מהחלטה כזו. בפרק זה אנו מציעים שיטה לאמוד את אי הוודאות בתהליך הרוויזיה (הערכה מחדש) של נתונים. אנו משתמשים באומדן זה כדי ליצור מרווח-חיזוי אסימטרי של הרוויזיה שתתאים לתקופת הפעילות שמפורסמת בהווה בזמן אמת. הנחת העבודה היא שהמרווח הזה הוא פועל יוצא ותלוי באורך הסדרה של כל פרק זמן, כלומר מספר הרוויזיות, בתוך הבציר הנוכחי, כחלק קצב צמיחת של תהליך. הגישה שלנו מרחיבה את ההנחה הרווחת בספרות שתהליך רוויזיות הוא הומוגני ולא עירוב של מספר תהליכים שונים. אנו מניחים שתהליך רוויזיות הוא פועל יוצא ותלוי במצב של תהליך הצמיחה, ולכן יש צורך למדל את התהליך באמצעות מודלים מרקוביים מוסתרים (Hidden Markov Models), ולאמוד את הפרמטרים עבור כל מצב. מתודולוגיה זו נבחנת כאן על בסיס נתונים היסטוריים של התוצר המקומי הגולמי (תמ"ג) של ישראל. אנו מראים כי קיימים הבדלים מובהקים בגודל ובסימן של הרוויזיות, התלויים במצב של הקצב הצמיחה. ליתר דיוק, הפרסום הראשוני נאמדים במצב של קצב צמיחה נמוך, קיימת תחילה הערכת יתר של הצמיחה, אבל פרסומים עוקבים אכן יראו ירידה נוספת של קצב הצמיחה. לעומת זאת כאשר הפרסום הראשוני נוצר במצב של קצב צמיחה גבוה, קיימת הערכת חסר או תת-הערכה ופרסומים עוקבים אכן יראו הגדלה של קצב הצמיחה.

פרק 4 דן בכיול וסיווג של מודלים "מעורבים" ליניאריים (linear mixed models). מבני נתונים אלה היו במקור ומראש מעבר לתחום המחקר הראשוני של מודל לינארי מוכללים, כמו במשפחות של מודלים ניבחרים מסוג L1 ו-L2. בקרה ושליטה של המודלים עם "השפעות המעורבות" מאפשרת לחוקרים את הגמישות לבנות מודלים מורכבים יותר של נתונים עם השפעות אקראיות ברמת הפרט (subject specific). מבני נתונים מסוג זה נפוצים בתחומים יישומיים של מחקרים שונים, ומאפשרים לקלוט את המורכבות שנצפתה בעולם האמיתי. פרק זה מגדיר הרחבה לקבוצת הקנסות הנוכחית שנחקרה במודלים מעורבים ליניאריים ובמודלים מעורבים ליניאריים מוכללים. הקנס נקרא Linear Mixed Model Elastic Net (LMMEN). סיבת או מטרת הקנס היא לבחור בו זמנית גם את ההשפעות הקבועות וגם האקראיות במודל, תוך מתן רמות גבוהות של קורלציה הנמצאים בתוך שני סוגי ההשפעה. התוצאות התיאורטיות והסימולציות המשוות קנסות שונות ומתחרות, גם הן נדונות בפרק. חישוב יעיל ומדויק של אגירת "חכמת ההמון", על מנת לענות על שאלות הסתברותיות שמעצבות מדיניות כלשהי, היא תמיד

## תקציר

האתגר של קבלת החלטות בעידן של מאגרי מידע רב-מימדיים הופך ליותר ויותר קשה עם השיפור ביכולת לאחסן מידע שפרחה בשני העשורים האחרונים. הבעייה או ההתלבטות לגבי "סיבה ומסובב" כאן, קיימת כאשר מנסים לברר איזה שאלות בקביעת מדיניות מחייבות התקדמות בשיטות מחקר של הסקה וניבוי סטטיסטיים, כאשר מצד שני שיפור בשיטות מחקר וניבוי טוב יותר יכולים לפתוח ערוצי תחקיר חדשים שיביאו לשיפור מדיניות. עבודת-תזה זו בוחנת כיצד התקדמויות בנייתוח נתונים רב-מימדיים וחילוץ אוטומטי, יכולים להשפיע על הסקה סטטיסטית ואופן קבלת ההחלטות במדיניות, בסביבה עם לחץ זמן ומוגבלות זמן. עבודת-תזה זו היא בעלת שלושה פרקים עיקריים שעוסקים בהיבטים הטכניים, התאורטיים והישומיים בסטטיסטיקה המיועדים לפתור בעיות מדיניות עם משמעות מעשית עכשווית.

לאחר פרק ההקדמה, המציגה הקשר לכל אחד מהפרקים, נציג את פרק 2 עוסק בבעית "ניבוי ההווה" – Nowcasting. מושג "ניבוי הווה" אינו מובן מאליו, כיוון שהוא מתייחס להווה במקום העתיד – שכן אנו תמיד בהווה, אך יש הרבה החלטות בקביעת מדיניות שתלויות בזמינות המידע העכשווי בעת ההחלטה. יתרה מכך, מועד סופי שנקבע מראש מחייב קבלת החלטה בלתי תלוי בזמינות המידע שכבר נצבר או ימשיך להצטבר. ישנם מספר תחומים אשר בהם חייבים לקבוע ולהעריך את המצב העכשווי של המערכת, כאשר המערכת עדיין ממשיכה לצבור מידע – כגון מערכות כלכלה, אפידמיולוגיה, מימון, מטאורולוגיה, וטכנולוגיות עתירות-ידע העוסקות באיסוף ידע מרשתות חברתיות. ככל שטכנולוגיה משפרת את יכולתה למדוד ולאחסן כמויות גדולות יותר של נתונים, הרי ההגדרה של ההווה, כלומר 'עכשיו' יכול להיות הגדרה מפורטת יותר היוורדת ליותר פרטים. בעיה זו בולטת יותר כשהמערכת המנבאת נותנת תחזית המבוססת על נתונים רב-מימדיים, כאשר מספר המשתנים גדול ממספר התצפיות. במקרה זה הפתרון המקובל של ריבועים פחותים לא אפשרי, ומחייב הורדת מימד של הבעייה. בפרק זה אנו בוחנים מספר שיטות להורדת מימד, ומשווים בין שני כיווני מחשבה שימושיים מרכזיים – כאשר ההורדה המימדית מותנית וגם בלתי מותנית במשתנה התלוי. לאורך הפרק, ההשוואה התיאורטית ניתמכת על ידי ונעזרת בדוגמאות מהאתגר של יצירת מדיניות לקביעת הריבית במשק הישראלי על ידי בנק ישראל.

פרק 3 עוסק בפן אחר של הצטברות רציפה ומתמשכת של מידע. אנו בוחנים סביבת נתונים המתעדכנת באופן מתמיד, ובכך יוצרים סביבה שבה אין ערך סופי ברור של תהליך מבוקש. מצב זה יוצר שכבה

**עבודה זו נעשתה בהדרכתו**

**של פרופ' יעקב ריטוב**

# סיווג וניטור תהליכים ברב מימד

חיבור לשם קבלת תואר דוקטור לפילוסופיה

מאת

יונתן סידי

הוגש לסנט האוניברסיטה העברית

סיון תשע"ח