Essays on Forecast Evaluation and Financial Econometrics

Kasper Lund-Jensen

Nuffield College & University of Oxford

Trinity Term 2013

Essays on Forecast Evaluation and Financial Econometrics

Kasper Lund-Jensen
Nuffield College & University of Oxford

Thesis submitted for the Degree of Doctor of Philosophy at the
Department of Economics, University of Oxford, Trinity Term 2013.

Abstract

This thesis consists of three papers that makes independent contributions to the fields of forecast evaluation and financial econometrics. As such, the papers, chapter 1-3, can be read independently of each other.

In Chapter 1, "Inferring an agent's loss function based on a term structure of forecasts", we provide conditions for identification, estimation and inference of an agent's loss function based on an observed term structure of point forecasts. The loss function specification is flexible as we allow the preferences to be both asymmetric and to vary non-linearly across the forecast horizon. In addition, we introduce a novel forecast rationality test based on the estimated loss function. We employ the approach to analyse the U.S. Government's preferences over budget surplus forecast errors. Interestingly, we find that it is relatively more costly for the government to underestimate the budget surplus and that this asymmetry is stronger at long forecast horizons.

In Chapter 2, "Monitoring Systemic Risk", we define systemic risk as the conditional probability of a systemic banking crisis. This conditional probability is modelled in a fixed effect binary response panel-model framework that allows for cross-sectional dependence (e.g. due to contagion effects). In the empirical application we identify several risk factors and it is shown that the level of systemic risk contains a predictable component which varies through time. Furthermore, we illustrate how the forecasts of systemic risk map into dynamic policy thresholds in this framework. Finally, by conducting a pseudo out-of-sample exercise we find that the systemic risk estimates provided reliable early-warning signals ahead of the recent financial crisis for several economies.

Finally, in Chapter 3, "Equity Premium Predictability", we reassess the evidence of out-of-sample equity premium predictability. The empirical finance literature has identified several financial variables that appear to predict the equity premium in-sample. However, Welch & Goyal (2008) find that none of these variables have any predictive power out-of-sample. We show that the equity premium is predictable out-of-sample once you impose certain shrinkage restrictions on the model parameters. The approach is motivated by the observation that many of the proposed financial variables can be characterised as 'weak predictors' and this suggest that a James-Stein type estimator will provide a substantial risk reduction. The out-of-sample explanatory power is small, but we show that it is, in fact,economically meaningful to an investor with time-invariant risk aversion. Using a shrinkage decomposition we also show that standard combination forecast techniques tends to 'overshrink' the model parameters leading to suboptimal model forecasts.

Acknowledgements

It is a great pleasure to have the opportunity to thank my supervisors Kevin Sheppard and Neil Shephard for their support, encouragement and guidance. During my time as a doctoral student they always had time to meet with me and discuss new ideas or results, and their sharp and stimulating supervision has been invaluable.

Several other people have contributed directly or indirectly to the writing of this thesis. In particular, I would like to thank Steve Bond, Martin Browning, John Campbell, Jennie Castle, David Hendry, Ian Jewitt, Shin Kanya, Jeremy Large, Steve Nickell, Bent Nielsen, Heino Bohn Nielsen, Diaa Noureldin, Cavit Pakel, Andrew Patton, Anders Rahbek, Gerhard Toews, Darren Upton and Louise Willerslev-Olsen.

I would also like to thank the International Monetary Fund and also to thank, personally, Dimitri Demekas, Nicolas Blancher, Srobona Mitra and Laura Kodres. I had the pleasure of visiting the Fund during 2011 where I contributed to the Global Financial Stability Report (GFSR), September 2011. More importantly, the idea for the second chapter of this thesis was conceived during my visit, and the dataset used in the paper is from the GFSR report.

My passion for economics and econometrics was ignited by the BSc in Economics programme at the University of Copenhagen. I would like to thank all my teachers in the economics department and my classmate Louise Willerslev-Olsen for teaching me all the basic economic principles upon which this thesis is based.

There is a bit more to life than econometrics, and I must also thank my good friends Ana, Andrey, Andreas, Arnaud, Cavit, Diaa, Gerhard, Kadu, Lawrence and Mathias for their friendship which made life a lot more pleasant during my studies in Oxford. I also thank Anthony Harris, Carlos Sancho and the rest of the Nuffield Lions for a lot of fun games on the football pitch.

During my doctoral studies at the University of Oxford I have been extremely priviliged to receive financial support from several different sources. Financial support from the Scatcherd European Scholarship, the Economic and Social Research Council (ESRC), Nuffield College, the David Walton Memorial Fund, Oxford-Man Institute of Quantitative Finance and the Department of Economics at the University of Oxford are gratefully acknowledged.

Finally, I am especially grateful to my gorgeous wife Liliane and my parents, Per and Anne Dorthe, whose love and encouragement has been criticial for the writing of this thesis. Without them this project would never have been completed and I dedicate this thesis to them and to my wonderful daughter Sophia.

Kasper Lund-Jensen

Oxford, May 2013

Word Count: The body of this thesis consists of 172 pages with a typical page containing 300 words. Therefore, this thesis is approximately 51,600 words long. This thesis was typed using LyX, which was also used to create all tables. Matlab was used to conduct all simulations and estimations.

Contents

List of Figures

List of Tables

Chapter 1

Inferring an agent's loss function based on a term structure of forecasts

Economic theory is often based on the assumption that agents are rational when they form forecasts of economic variables. It is challenging to test this assumption empirically as an agent's loss function typically is unobserved. In this paper we provide conditions for identification, estimation and inference of an agent's loss function based on an observed term structure of point forecasts. The loss function specification is flexible as we allow the preferences to be both asymmetric and to vary non-linearly across the forecast horizon. In addition, we introduce a novel forecast rationality test based on the estimated loss function. We employ the approach to analyse the U.S. Government's preferences over budget surplus forecast errors. Interestingly, we find that it is relatively more costly for the government to underestimate the budget surplus and this asymmetry is stronger at long forecast horizons. This may reflect a political bias that arises because it is politically costly to cut government expenses, or to raise taxes, and that this effect is stronger at longer forecast horizons due to uncertainty related to elections.

Keywords: Forecast Rationality, Loss Function, Term Structure of Forecasts.
Journal of Economic Literature Codes: C12; C13; C32.

1.1 Introduction

Economic theory is often based on the assumption that agents are rational when they form forecasts of economic variables. It is therefore of great interest to test this forecast rationality assumption empirically. Unfortunately, this is a challenging task as the agent's loss function is typically unobservable. The literature has traditionally sidestepped this issue by assuming that the agent's preferences over forecasting errors can be described by mean squared error loss (e.g. Mincer & Zarnowitz 1969). This is problematic as there are often good reasons to believe that the loss function depends asymmetrically on positive and negative forecast errors and potentially on other factors. This issue has been discussed by, among others, Granger (1969), Varian (1975), Zellner (1986), Granger & Newbold (1986), Christoffersen & Diebold (1997), Granger & Pesaran (2000) and Granger & Machina (2006). Therefore, rather than postulate that the agent has a particular loss function, a potentially superior approach is to attempt to estimate the agent's loss function and to utilize this information in a forecast rationality test. Importantly, the estimated loss function is also an interesting object in its own right, as it provides information about the agent's objectives and motives.

In this paper, we provide conditions for identification, estimation and inference of an agent's loss function based on an observed term structure of point forecasts. We also propose a novel forecast rationality test based on the estimated loss function. The key advantage of this approach is that we utilize the information in the complete term structure of forecasts. This allows for a flexible loss function specification and we are able to identify more features of the agent's loss function. Simulation results also suggest that the rationality test has good finite sample properties.

The methodology developed in this paper is motivated by the fact that forecasts recorded at multiple horizons, e.g. from one to several years into the future, are frequently observed. For example, the forecasts produced by the IMF (World Economic Outlook), the Bank of England, the Board of the Federal Reserve and the U.S. Office of Management and Budget all cover multiple horizons. With the availability of such multi-horizon forecasts, it is natural to develop new methods that exploit the information in the entire term structure of forecasts. By simultaneously exploiting information across several horizons, rather than focusing separately on individual horizons, multi-horizon forecast tests offer the potential of obtaining more precise estimates of the agent's unobserved loss function and of drawing more powerful conclusions about the agent's ability to form optimal (rational) forecasts. In addition, as we will discuss in more detail in section 1.3, one is able to identify more features of an agent's loss function based on a term structure of forecasts.

In the empirical section, we employ the approach to analyse the U.S. Administration's preferences over budget surplus forecast errors. Interestingly, the empirical analysis suggests that the U.S. Administration's loss function is asymmetric and that this asymmetry is stronger at long forecast horizons. We conjecture that this loss function structure may arise as a result of political pressure and uncertainty related to political elections.

The rest of the paper is organised as follows. Section 1.2 contains an overview of the related literature. Section 1.3 describes the modeling framework and presents the main theoretical results. Section 1.4 presents some Monte Carlo simulation results that analyse the finite sample properties of the estimator of the loss function's shape parameters and the forecast rationality test. In section 1.5, we use the methodology to analyse the U.S. Government's preferences over budget surplus forecast errors. Finally, section 1.6 concludes.

1.2 Literature Overview

This paper contributes to the theoretical literature on loss function identification and forecast rationality testing. Economic theory is often based on a forecast rationality assumption and considerable effort has been devoted to empirically assessing the validity of this assumption. Empirical work has traditionally relied on the additional assumption that the agent's loss function, $\mathcal{L}(\cdot)$, can be approximated by mean squared error (MSE) loss:

$$\mathcal{L}(Y_{t+h}, \hat{Y}_{t+h,t})_{MSE} = (Y_{t+h} - \hat{Y}_{t+h,t})^2,$$

where t denotes the time at which the forecast is formed, h denotes the forecast horizon, Y_{t+h} is the economic variable of interest and $\hat{Y}_{t+h,t}$ is the forecast based on information available at time t. Under covariance stationarity, optimal MSE forecasts satisfy the following conditions (see, e.g. Diebold & Lopez 1996 or Patton & Timmermann 2007*a*):

1. The optimal forecast, $\hat{Y}^{*MSE}_{t+h,t}$, is given by the conditional expectation,

$$\hat{Y}^{*MSE}_{t+h,t} = \mathbb{E}[Y_{t+h}|\mathcal{F}_t]\,,$$

   and is thus conditionally (and unconditionally) unbiased.

2. The h-step ahead forecast error exhibits zero serial covariance beyond lag $(h-1)$:

$$\mathbb{E}[(Y_{t+h} - \hat{Y}^{*MSE}_{t+h,t})(Y_{t+h-j} - \hat{Y}^{*MSE}_{t+h,t})] = 0 \ for \ j > h-1.$$

3. The unconditional variance of the forecast error is a weakly increasing function of the forecast horizon:

$$\mathbb{E}[(Y_{t,t-h} - \hat{Y}^{*MSE}_{t,t-h})^2] \leq \mathbb{E}[(Y_{t,t-h-j} - \hat{Y}^{*MSE}_{t,t-h-j})^2] \ for \ j \geq 1.$$

Therefore, a joint test of forecast rationality and MSE loss can be conducted by empirically assessing the validity of any of these three conditions. A classic example is the popular Mincer-Zarnowitz regression (Mincer & Zarnowitz 1969), $Y_{t+h} = \alpha + \beta\hat{Y}_{t+h,t} + u_{t+h}$, which tests whether the realizations and the forecasts fall around the 45-degree line, i.e. whether $\alpha = 0$ and $\beta = 1$ (*property 1*). A more recent example is Patton & Timmermann (2012) who propose a new forecast rationality test, under MSE preferences, which evaluates whether the variance of the forecasts is a weakly increasing function in the forecast horizon (*property 3*).[1] A drawback of these tests is that they test the joint hypothesis of forecast rationality and MSE loss and therefore impose strong parametric assumptions on the form of the agent's unobservable loss function. For example, Elliott, Komunjer & Timmermann (2008) show that the Mincer-Zarnowitz test can lead to false rejections of forecast rationality when the

[1] Additionally Patton & Timmermann (2012) derive a modified Mincer-Zarnowitz-type test incorporating multi-horizon forecasts

loss function is asymmetric. This is an important observation as there are often good reasons to believe that the loss function depends asymmetrically on positive and negative forecast errors. This point has been discussed by, among others, Granger (1969) Varian (1975), Zellner (1986), Granger & Newbold (1986), Christoffersen & Diebold (1997), Granger & Pesaran (2000) and Granger & Machina (2006). In chapter 3 of this thesis, "Equity Premium Predictability: Weak Predictors, Shrinkage and Economic Value", we also show that an investor's preferences over volatility forecast errors can be asymmetric.

More recently, Elliott, Komunjer & Timmermann (2005) and Patton & Timmermann (2007*b*) have developed new tests of forecast rationality that relax the assumption of mean squared error loss. Interestingly, based on two different applications, both papers reject the joint hypothesis of forecast rationality and mean squared error loss but find that forecast rationality cannot be rejected when allowing for a more general loss function. These findings provide empirical evidence for the view that mean squared error loss does not necessarily provide an accurate description of an agent's preferences over the forecast error. Komunjer & Owyang (2012) propose a multivariate forecast rationality test that extends the approach proposed by Elliott, Komunjer & Timmermann (2005).

Though the practice of making multi-horizon forecasts is relatively widespread this feature is seldom incorporated in rationality testing. The exceptions are Patton & Timmermann (2012) and Capistran (2007). Capistran (2007) generalizes the property of non-decreasing variance of forecast errors when loss is MSE (*property 3*) to a property of non-decreasing expected loss when the loss function differs from MSE. The resulting tests are complimentary to the rationality tests in Elliott, Komunjer & Timmermann (2005) and Patton & Timmermann (2007*b*).

The main theoretical contribution of this paper consist of developing a model framework for identifying and estimating an agent's loss function and use this information to construct a flexible rationality test. The approach can be thought of as a generalization of Elliott, Komunjer & Timmermann (2005) to the case where one observes a term structure of point forecasts and we argue that this approach has several advantages. Firstly, the approach allows for a more flexible loss function specification and enables us to identify more features of the agent's loss function. More specifically, the agent's preferences are allowed to vary non-linearly across the forecast horizon. Secondly, if the agent's preferences depend on the forecast horizon, classic forecast rationality tests tend to falsely reject forecast rationality. In contrast, our forecast rationality test has good size properties under the null where the agent's loss function depends on the forecast horizon. This is illustrated in a simulation experiment. The proposed model framework is motivated by the fact that forecasts are often recorded at multiple horizons, e.g. from one year to several years into the future. For example, the forecasts produced by the IMF (World Economic Outlook), the Bank of England, the Board of the Federal Reserve and the U.S. Office for Budget and Management all cover multiple horizons.

1.2.1 Budget Surplus Forecasts and the Political Business Cycle

This paper also contributes to the empirical literature by employing the proposed novel methodology to analyse the U.S. administration's preferences over budget surplus forecast errors. The idea behind this application is to evaluate whether we can reject that political pressure impacts the administration's budget projections.

The empirical application is related to the literature on political business cycles. In the original political business cycle models in the style of Nordhaus (1975) and MacRae (1977) politicians care about winning elections, either for the sake of power itself or to pursue partisan policies once elected. The assumption is that voters reward politicians who instate low taxes, offer high expenditures and guarantee low unemployment. Politicians in office thus face an incentive to stimulate aggregate demand and reduce unemployment by increasing public spending prior to elections. This opportunistic policy raises re-election chances when voters are myopic. But increasing public spending may be difficult to justify in a situation where deficits are high or may even be impossible if legislative caps and constraints on the budget deficit or on public debt prevent deficits over a certain threshold. One way to overcome these spending barriers and increase fiscal latitude is to produce deliberately optimistic forecasts of the future budget deficit. Likewise an optimistic forecast of the budget deficit can be used to postpone necessary structural reforms that would be unpopular with voters until after the election has been passed.

Rogoff & Sibert (1988) and Rogoff (1990) discard the assumption about myopic voters who favor politicians with high public spending. Instead voters are assumed to be rational but to have imperfect information about the competence of politicians, where competence is defined as the ability to deliver public goods for a given level of tax revenue. Due to imperfect information voters cannot determine whether high expenditures reflect politicians' true competence or is simply due to temporarily inflated expenditures that will result in budgetary cuts later on. In this version of the political business cycle model some incumbents choose to increase public spending before an election to signal high competence and thereby increase chances of reelection. Shi & Svensson (2006) consider a similar version of the political business cycle model but change the signaling problem to a moral hazard problem by assuming that politicians do not know their own type. The implication is that all incumbents regardless of competence will face an incentive to boost public spending in a bid/attempt to appear competent. In recent work Bohn (2011) explicitly incorporates the role of budget deficit forecasting into these latter types of political business cycle models. In Bohn (2011) the intuitive argument is that even though forecast manipulations lead to reputational costs similar to Alesina, Roubini & Cohen (1997) if the government can appear more competent by raising the level of transfers, then it should appear even more competent, if it can convince (at least some) voters of being able to do so without incurring large deficits. Bohn (2011) thus incorporates the budget deficit forecast as a variable that the incumbent can use to increase reelection chances. Though no closed form solution for the forecast is found, these models are a first step towards developing models of optimal forecasting behavior. In a different strand of the political business cycle literature Hibbs Jr (1977) and Alesina (1987) stress that political parties are motivated by ideology rather than simply wanting to stay in power. In these models political business cycles occur because of partisan effects. Right-wing governments are assumed to be inflation adverse while left-wing governments are unemployment adverse. As argued by Brück & Stephan (2006) a left-wing government may therefore be too optimistic in its unemployment forecast which translates into an overoptimistic forecast of tax revenues and budget deficit. Conversely right-wing politicians may want to pursue low inflation and to reduce the size of government. As argued by Boylan (2008) Republican Administrations may in this case ”overestimate inflation and underestimate government revenues in order to strengthen the case for limiting monetary and spending growth”.

Empirically the forecast performance of the US Administration has recently been evaluated by Frankel (2011), Frankel & Schreger (2012) and Kliesen & Thornton (2012). They all note that the bias in budget forecasts display a tendency to increase with the forecast horizon. One advantage of the methodology developed in this paper, is that it allows us to identify whether the increasing bias structure is simply a result of higher forecast error variance at long forecast horizons or whether it is due to more fundamental forecast-horizon varying features of the loss function.

1.3 Identification and Estimation of Loss Function

In this section we define the modeling framework and present the main theoretical results. More specifically, we study the forecaster's optimal problem and establish conditions under which the parameters of the loss function are identified based on an observed term structure of point forecasts. We also provide conditions for consistency and asymptotic normality of the loss function's shape parameters. Finally, we introduce a novel forecast rationality test that allows for a general class of loss functions and exploits the information in the entire term structure of point forecasts.

1.3.1 Model Setup

Our setup is as follows: let $\mathbf{X} \equiv \{\mathbf{X}_t : \Omega \to \mathbb{R}^{\mathrm{m}+1}, \mathrm{m} \in \mathbb{N}^*, \mathrm{t} = 1, ..., \mathrm{T}+\mathrm{H}\}$ be a stochastic process defined on a complete probability space $(\Omega,\ \mathcal{F},\ Pr)$, where $\mathcal{F} = \{\mathcal{F}_t, t = 1, ..., T + H\}$

and $\mathcal{F}_t$ is the $\sigma - field$, $\mathcal{F}_t \equiv \sigma\{\mathbf{X}_s,\, s \leq t\}$. Let $\mathbf{X}_t = (Y_t, \mathbf{V}_t)$ where $Y_t \in \mathbb{R}$ denotes the economic variable of interest and $\mathbf{V}_t \in \mathbb{R}^{\mathrm{m}}$ denotes a $m \times 1$ vector of time t measurable random variables known to the agent. The agent's problem is to forecast the path of Y_t *H-periods* ahead, $\mathbf{Y}_{t+H} = (Y_{t+1}, Y_{t+2}, ..., Y_{t+H})^T$, conditional on the information set $\mathcal{F}_t$. The term structure of point forecasts, conditional on $\mathcal{F}_t$, is denoted by
$\hat{\mathbf{Y}}_{t+H,t} = (\hat{Y}_{t+1,t}, ..., \hat{Y}_{t+H,t})^T$ and belong to $\mathcal{Y}$ a subset of $\mathbb{R}^{\mathrm{H}}$. Generally, an agent's optimal forecast, $\mathbf{Y}^*_{t+H,t} \in \mathcal{Y}$, depends on the relative costs of negative and positive forecast errors and possibly other factors. We will assume that the agent's preferences can be described by a loss function, $\mathcal{L} : \mathcal{Y} \times \mathbb{R}^{\mathrm{H}} \rightarrow \mathbb{R}$, that maps the forecasts and realizations into the real line. The first main assumption refers to how the agent forms optimal (rational) forecasts.

Assumption 1 (Rationality). *The agent constructs rational (optimal) forecasts of* $\mathbf{Y}_{t+H}$*, by solving*

$$\mathbf{Y}^*_{t+H,t} = \underset{\hat{\mathbf{Y}}_{t+H,t} \in \mathcal{Y}}{arg\,min}\ \mathbb{E}[\mathcal{L}(\mathbf{Y}_{t+H}, \hat{\mathbf{Y}}_{t+H,t}; \boldsymbol{\alpha}_0)|\mathcal{F}_t], \tag{1.3.1}$$

where $\mathcal{F}_t \equiv \sigma\{\mathbf{X}_s,\, s \leq t\}$ *is the* $\sigma - algebra$ *that represents the agent's information set and* $\mathcal{L}(\cdot; \boldsymbol{\alpha}_0)$ *is the agent's loss function.*

This assumption states that the forecaster is rational, or is an expected loss minimiser, and economic theory is often based on it. It is of great interest to test this assumption empirically but this is complicated by the fact that the agent's loss function generally is unobservable. Classic tests of forecast rationality (e.g. Mincer & Zarnowitz 1969) impose strong parametric assumptions on the loss function by assuming that it can be approximated by mean squared error. This paper relaxes this assumption by assuming that the loss function belongs to a flexible family of functions.

Assumption 2 (Loss Function Specification). *The agent's loss function has the following form*

$$\mathcal{L}(\mathbf{Y}_{t+H}, \hat{\mathbf{Y}}_{t+H,t}; \boldsymbol{\alpha}_0) = \sum_{h=1}^{H} [f_h(\boldsymbol{\alpha}_0) + (1 - 2f_h(\boldsymbol{\alpha}_0)) \cdot 1\{\hat{\epsilon}_{t+h,t} < 0\}] |\hat{\epsilon}_{t+h,t}|^{\lambda},$$

where $\hat{\epsilon}_{t+h,t} = Y_{t+h} - \hat{Y}_{t+h,t}$, $\lambda \in \mathbb{R}^+$, $\boldsymbol{\alpha}_0 \in \Theta \subset \mathbb{R}^k$ is a $k \times 1$ vector of unknown parameters and $\mathbf{f}(\boldsymbol{\alpha}_0) = (f_1(\boldsymbol{\alpha}_0), ..., f_H(\boldsymbol{\alpha}_0))^T$ is a measurable function, $\mathbf{f} : \Theta \to (0,1)^H$.

This time separable class of loss functions nests several popular loss functions: (i) squared loss for $\lambda = 2$ and $\mathbf{f}(\boldsymbol{\alpha}_0) = 0.5\boldsymbol{\iota}$, (ii) absolute deviation loss for $\lambda = 1$ and $\mathbf{f}(\boldsymbol{\alpha}_0) = 0.5\boldsymbol{\iota}$, (iii) quad-quad loss for $\lambda = 2$ and $\mathbf{f}(\boldsymbol{\alpha}_0) = \alpha_0\boldsymbol{\iota}$ and (iv) lin-lin (or "tick function") loss for λ=1 and $\mathbf{f}(\boldsymbol{\alpha}_0) = \alpha_0\boldsymbol{\iota}$. More importantly, this class of loss functions also nests more general loss functions that allow the forecaster's preferences to vary non-linearly across the forecast horizon, h, such that the relative costs of positive and negative forecast errors can depend on the forecast horizon.

When inferring an agents loss function, we will take λ and the functional form of $\mathbf{f}(\boldsymbol{\alpha}_0)$ as given and estimate the parameter vector, $\boldsymbol{\alpha}_0 \in \Theta \subset \mathbb{R}^k$. While this approach imposes some degree of structure on the loss function specification, it still allows for a rich family of loss functions.

1.3.2 Identification of the Loss Function's Shape Parameters

As we mentioned in the introduction, one of the key challenges when testing forecast rationality is that the agent's loss function is unobserved. In this paper we deal with this issue by estimating the loss function based on an observed term structure of point forecasts. The idea is to derive a set of necessary and sufficient conditions of forecast optimality and use these to identify the shape parameters of the loss function. More specifically, under Assumption 1 and 2, and some additional technical assumptions provided in appendix A (Assumption 3), we are able to derive the following necessary and sufficient conditions for forecast optimality (or rationality).

Proposition 1 (Moment Conditions). *Under Assumption 1-3, and for $\boldsymbol{\alpha}_0 \in \Theta$, the forecasts $\mathbf{Y}^*_{t+H,t}$ are optimal (rational) if and only if*

$$\mathbb{E}[g(\mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t}; \boldsymbol{\alpha}_0)|\mathcal{F}_t] = 0 \Leftrightarrow \mathbf{Y}^*_{t+H,t} \text{ is optimal} \tag{1.3.2}$$

where

$$g(\mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t}; \boldsymbol{\alpha}_0) = diag(|\mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t}|^{\lambda-1})(1\{\mathbf{Y}_{t+H} - \hat{\mathbf{Y}}^*_{t+H,t} < 0\} - \mathbf{f}(\boldsymbol{\alpha}_0))$$

Proof. See appendix B. □

Proposition 1 shows that under fairly weak assumptions, the sequence of optimal point forecast paths satisfy a set of conditional moment conditions. Therefore, all time t measurable variables are exogenous instruments:
$\mathbb{E}[\mathbf{Z}_t \otimes g(\mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t}; \boldsymbol{\alpha}_0)] = 0, \forall \mathbf{Z}_t \in \mathcal{F}_t$. This is a useful result and can be used to identify $\boldsymbol{\alpha}_0$. The following proposition formalises this idea.

Proposition 2 (Identification). *Suppose that Assumption 1-3 hold and let*

$$Q(\boldsymbol{\alpha}) = \mathbb{E}[\mathbf{Z}_t \otimes g(\mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t}; \boldsymbol{\alpha}_0)]^T \mathbf{W} \mathbb{E}[\mathbf{Z}_t \otimes g(\mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t}; \boldsymbol{\alpha}_0)]$$

where $\mathbf{W}$ *is a* $nH \times nH$ *positive definite weight matrix,* $\mathbf{Z}_{\mathrm{t}} \in \mathcal{F}_t \equiv \sigma\{\mathbf{X}_s,\ s \leq t\}$ *is a* $n \times 1$ *(*$n \leq m$*) vector and* $g(\cdot)$ *is defined as in Proposition 1. Then* $\boldsymbol{\alpha}_0$ *is the unique minimum of* $Q(\boldsymbol{\alpha})$ *on* Θ*.*

Proof. See appendix B. □

Proposition 2 shows that the true value $\boldsymbol{\alpha}_0$ is globally identified as the unique minimum of the objective function $Q(\boldsymbol{\alpha})$. Note that in order to identify $\boldsymbol{\alpha}_0$ the forecast user is not required to observe the full vector of variables used by the forecaster,
$\mathcal{F}_t = \sigma\{(Y_s, \mathbf{V}_s),\ s \leq t\}$, but only a subvector of these variables,
$\mathcal{F}^Z_t = \sigma\{(Y_s, \mathbf{Z}_s),\ s \leq t\} \subset \mathcal{F}_t$. This is a powerful result as we would generally expect the forecaster to have access to private information which is outside of the econometrician's information set. In addition, it is also worth pointing out that, contrary to Elliott, Komunjer & Timmermann (2005), we are able to identify a shape parameter that allows the preferences to depend on the forecast horizon.

Although Proposition 2 states that $\boldsymbol{\alpha}_0$ is (globally) identified, there does not exist a closed form solution of $\boldsymbol{\alpha}_0$ for a general non-linear specification of $\mathbf{f}(\boldsymbol{\alpha})$. However, it is possible to obtain a closed form solution for $\boldsymbol{\alpha}_0$ if one restricts $\mathbf{f}(\boldsymbol{\alpha})$ to have a particular form. The following lemma formalises this.

Lemma 3. *Suppose that the assumptions of Proposition 2 hold, $\lambda \in \mathbb{R}^+$ and $\mathbf{f} : \Theta \to (0,1)^H$. Furthermore, suppose that $\mathbf{f}(\boldsymbol{\alpha}) = \boldsymbol{\Pi}\boldsymbol{\alpha}$ where $rank[\boldsymbol{\Pi}] = k \leq H$, then*

$$\begin{aligned}\boldsymbol{\alpha}_0 &= \left(\mathbb{E}[\mathbf{Z}_t \otimes diag(|\mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t}|^{\lambda-1})\Pi)]^T \mathbf{W}\mathbb{E}[\mathbf{Z}_t \otimes diag(|\mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t}|^{\lambda-1})\Pi]\right)^{-1} \\ &\quad \times \mathbb{E}[\mathbf{Z}_t \otimes diag(|\mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t}|^{\lambda-1})\Pi)]^T \mathbf{W}\mathbb{E}[\mathbf{Z}_t \otimes diag(|\mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t}|^{\lambda-1})1\{\mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t} < 0\}]\end{aligned}$$

where $\mathbf{W}$ is a positive definite weight matrix.

From a computational point of view, this is a very useful result as the closed form solution implies that there is no need for solving a non-linear minimization problem numerically. Having provided conditions for identification, we now turn our attention to the estimation of the loss function's shape parameter, $\boldsymbol{\alpha}_0$.

1.3.3 Estimation of $\boldsymbol{\alpha}_0$

The identification result in Proposition 2 is the starting point of our estimation and rationality testing procedure. Having established that the parameter $\boldsymbol{\alpha}_0$ is globally identified we now turn to the problem of estimating $\boldsymbol{\alpha}_0$ by minimizing an empirical counterpart of $Q(\boldsymbol{\alpha})$.

Under Assumption 1 and 2, and some additional regularity conditions, we will in the following show that it is possible to consistently estimate $\boldsymbol{\alpha}_0$ based on a sequence of T realized forecast error paths $\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t}\}_{t=R+1}^{R+T}$, where $\hat{\boldsymbol{\epsilon}}^*_{t+H,t} = (\hat{\epsilon}^*_{t+1,t}, ..., \hat{\epsilon}^*_{t+H,t})^T$. Based on Proposition 2, we know that

$$\boldsymbol{\alpha}_0 = \underset{\boldsymbol{\alpha}\in\Theta}{arg\,min} \left\{\mathbb{E}[\mathbf{Z}_t \otimes g(\boldsymbol{\epsilon}^*_{t+H,t};\boldsymbol{\alpha})]^T \mathbf{W}\mathbb{E}[\mathbf{Z}_t \otimes g(\boldsymbol{\epsilon}^*_{t+H,t};\boldsymbol{\alpha})]\right\}$$

for any $Hn \times Hn$ positive definite matrix, $\mathbf{W}$. Therefore, asymptotically, the particular choice of weight matrix is irrelevant for consistency as long as it is positive definite. However, in order to minimise the variance of the estimator based on a finite sample, we choose a particular weight matrix when forming the estimator of $\boldsymbol{\alpha}_0$:

$$\hat{\boldsymbol{\alpha}}_T = \underset{\boldsymbol{\alpha}\in\Theta}{arg\,min}\left\{\hat{\mathbf{h}}_T^T(\boldsymbol{\alpha})\hat{\mathbf{S}}^{-1}\hat{\mathbf{h}}_T(\boldsymbol{\alpha})\right\} \tag{1.3.3}$$

where

$$\hat{\mathbf{h}}_T(\boldsymbol{\alpha}) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{Z}_t \otimes diag(|\mathbf{Y}_{t+H} - \hat{\mathbf{Y}}^*_{t+H,t}|^{\lambda-1})(1\{\mathbf{Y}_{t+H} - \hat{\mathbf{Y}}^*_{t+H,t} < 0\} - \mathbf{f}(\boldsymbol{\alpha}))$$

and

$$\hat{\mathbf{S}}(\hat{\boldsymbol{\alpha}}_T) = \hat{\Gamma}_0(\hat{\boldsymbol{\alpha}}_T) + \sum_{j=1}^{T-1}\omega_{i,T}(\hat{\Gamma}_j(\hat{\boldsymbol{\alpha}}_T) + \hat{\Gamma}_j(\hat{\boldsymbol{\alpha}}_T)^T)$$

where $\omega_{i,T}$ is a kernel (e.g. Bartlett following Newey and West (1987) or Quadratic Spectral following Andrews (1991)). The choice of weight matrix will become clear when we derive the asymptotic distribution of the estimator. In order to show that the estimator is consistent we need to assume some additional regularity conditions, such that a law of large numbers applies. These assumptions are summarised in Assumption 4 in appendix A. We are now in a position to state the following consistency proposition.

Proposition 4 (Consistency). *Let Assumption 1-4 hold. Then, given* $\lambda \in \mathbb{R}^+$ *and* $\mathbf{f} : \Theta \to (0,1)^H$*, we have that*

$$\hat{\boldsymbol{\alpha}}_T \xrightarrow{p} \boldsymbol{\alpha}_0 \text{ as } T \to \infty$$

where $\hat{\boldsymbol{\alpha}}_T$ *is defined as in equation* (1.3.3)*.*

Proof. See appendix B. □

Proposition 4 shows that under fairly weak conditions, we are able to consistently estimate the shape parameter of the agent's loss function based on an observed term structure of forecasts. This is a powerful result but it does not allow us to perform inference based on a finite sample. In order to establish asymptotic normality we need to assume some stronger mixing conditions such that a general CLT applies. These additional regularity conditions are summarized in Assumption 5 in Appendix A. Under these additional assumptions, we are able to state the following asymptotic normality result.

Proposition 5 (Asymptotic Normality). Let Assumption 1-5 hold. Then, given

$\lambda \in \mathbb{R}^+$ *and* $\mathbf{f} : \Theta \to (0,1)^H$, we have that

$$T^{1/2}(\hat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left(\mathbf{m}(\boldsymbol{\alpha}_0)^T \mathbf{S}(\boldsymbol{\alpha}_0)^{-1} \mathbf{m}(\boldsymbol{\alpha}_0)\right)^{-1}\right)$$

where $\mathbf{m}(\boldsymbol{\alpha}_0) = \mathbb{E}\left[\mathbf{Z}_t \otimes diag(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1}) \frac{\partial \mathbf{f}(\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}}\right]$ and $\mathbf{S}(\boldsymbol{\alpha}_0)$ is given by

$$\mathbf{S}(\boldsymbol{\alpha}_0) = \boldsymbol{\Gamma}_0(\boldsymbol{\alpha}_0) + \sum_{j=1}^{H-1} (\boldsymbol{\Gamma}_j(\boldsymbol{\alpha}_0)^* + \boldsymbol{\Gamma}_j(\boldsymbol{\alpha}_0)^{*T})$$

$$\boldsymbol{\Gamma}_0(\boldsymbol{\alpha}_0) = \mathbb{E}\left[\mathbf{Z}_t \mathbf{Z}_t^T \otimes \left(diag(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda_0-1})(1\{\boldsymbol{\epsilon}^*_{t+H,t} < 0\} - \mathbf{f}(\boldsymbol{\alpha}_0))\right) \left(diag(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda_0-1})(1\{\boldsymbol{\epsilon}^*_{t+H,t} < 0\} - \mathbf{f}(\boldsymbol{\alpha}_0))\right)^T\right]$$

$$\begin{aligned} \boldsymbol{\Gamma}_j(\boldsymbol{\alpha}_0)^* &= \mathbb{E}\Big[\mathbf{Z}_t \mathbf{Z}_{t+j}^T \otimes \Big\{\mathbf{1}^j_{H\times H} \left(diag(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda_0-1})(1\{\boldsymbol{\epsilon}^*_{t+H,t} < 0\} - \mathbf{f}(\boldsymbol{\alpha}_0))\right)\Big\} \times \\ &\quad \left(diag(|\boldsymbol{\epsilon}^*_{t+H+j,t+j}|^{\lambda_0-1})(1\{\boldsymbol{\epsilon}^*_{t+H+j,t+j} < 0\} - \mathbf{f}(\boldsymbol{\alpha}_0))\right)^T\Big] \end{aligned}$$

where $\mathbf{1}^j_{H\times H} = diag(\boldsymbol{\iota}^j_H)$, $\boldsymbol{\iota}^j_H = (0,0,...,1,1,...,1)^T$.

Proof. See appendix B. □

Remark 1: The first n rationality conditions are martingale difference sequences and are therefore serially uncorrelated through time. However, the other (H-1)n rationality conditions are not and this creates the particular autocorrelation terms, $\boldsymbol{\Gamma}_j(\boldsymbol{\alpha}_0)^*$.

Remark 2: If $H = 1$ *, the long-run variance simplifies to the variance expression in Elliott, Komunjer and Timmermann (2005)*

$$\mathbf{S} = \mathbb{E}\left[\mathbf{Z}_t \mathbf{Z}_t^T |\boldsymbol{\epsilon}^*_{t+1,t}|^{2\lambda-2} (1\{\boldsymbol{\epsilon}^*_{t+1,t} < 0\} - \alpha_0)^2\right]$$

Proposition 2, 4 and 5 provide conditions for identification, estimation and inference of the shape parameters of the loss function, $\boldsymbol{\alpha}_0$. The estimator of $\boldsymbol{\alpha}_0$ is based on the Hn moment conditions we derived in Proposition 1. Naturally, if the moment conditions are violated for some reason, e.g. if the rationality assumption is false, we cannot expect the estimator of $\boldsymbol{\alpha}_0$ to have any nice properties. If $Hn = k$ the estimator is just identified and we can not assess whether the moment conditions are violated as the k *empirical* moment conditions holds exactly by construction. However, if $Hn > k$ the estimator is over-identified and we are able to conduct an overidentification test based on Proposition 5. This overidentification test provides a joint test of forecast rationality (Assumption 1) and that the agent's forecast error preferences can be approximated by the flexible loss function (Assumption 2). This next Corollary presents the result.

Corollary 6 (Rationality Test). *Let the assumptions of Proposition 5 hold. Suppose that* $\boldsymbol{\alpha}$ *is overidentified,* $nH - k > 0$*, then a forecast rationality test can be conducted through the J-test statistic*

$$\begin{aligned} \hat{J}_T &= T\left(\frac{1}{T}\sum_{\text{i=1}}^{\text{T}}\mathbf{Z}_t \otimes g(\mathbf{Y}_{t+H} - \hat{\mathbf{Y}}^*_{t+H,t}; \hat{\boldsymbol{\alpha}}_T)\right)^T \\ &\quad \times\hat{\mathbf{S}}^{-1}\left(\frac{1}{T}\sum_{\text{i=1}}^{\text{T}}\mathbf{Z}_t \otimes g(\mathbf{Y}_{t+H} - \hat{\mathbf{Y}}^*_{t+H,t}; \hat{\boldsymbol{\alpha}}_T)\right) \sim \chi^2_{nH-k} \end{aligned}$$

where $\hat{\mathbf{S}}$ *is as defined in* (1.3.3).

Remark: The rationality test in Corollary 6 is an (important) bi-product of the estimation procedure.

Classic tests of forecast rationality test the joint hypothesis of forecast rationality and mean squared error loss. Therefore, one can falsely reject forecast rationality based on a classic forecast rationality test if MSE loss is a poor description of the forecaster's preferences. For example, Elliott, Komunjer & Timmermann (2008) showed that the Mincer-Zarnowitz forecast rationality test of single-horizon forecasts can lead to false rejections if the forecaster's preferences can be described by Quad-Quad. The test proposed in Corollary 6 relaxes the assumption of MSE loss by allowing the loss function to belong to a flexible class of loss functions such that the agent's preferences can be asymmetric and depend on the forecast horizon.

1.3.4 Loss Function Specifications

The theoretical framework developed in this paper is applicable to a wide range of loss functions as we have imposed few restrictions on λ and $\mathbf{f} : \Theta \to (0, 1)^H$. For example, the agent's loss function is allowed to depend asymmetrically on negative and positive forecast errors and this asymmetry is allowed to vary non-linearly over the forecast horizon. In the empirical application, we will work with the following two novel loss function specifications:

Definition 1: The LHVQ (**L**ogarithmic **H**orizon **V**arying **Q**uad-quad) Loss Function is defined as:

$$\mathcal{L}(\hat{\boldsymbol{\epsilon}}_{t+H,t};\boldsymbol{\alpha})_{LHVQ} = \sum_{h=1}^{H}[\alpha_1 + \alpha_2 log(h) + (1-2[\alpha_1 + \alpha_2 log(h)]) \cdot 1\{\hat{\epsilon}_{t+h,t} < 0\}]\hat{\epsilon}^2_{t+h,t}$$

where $(\alpha_1, \alpha_2) \in \Theta \subset \mathbb{R}^2$.

Definition 2: The LHVL (**L**ogarithmic **H**orizon **V**arying **L**in-lin) Loss Function is defined as:

$$\mathcal{L}(\hat{\boldsymbol{\epsilon}}_{t+H,t};\boldsymbol{\gamma})_{LHVL} = \sum_{h=1}^{H}[\gamma_1 + \gamma_2 log(h) + (1-2[\gamma_1 + \gamma_2 log(h)]) \cdot 1\{\hat{\epsilon}_{t+h,t} < 0\}]|\hat{\epsilon}_{t+h,t}|$$

where $(\gamma_1, \gamma_2) \in \Theta \subset \mathbb{R}^2$.

Note that the conditions for Lemma 3 are satisfied so there exists a closed form solution for the estimator of the loss function shape parameters under LHVQ or LHVL preferences.

The theoretical framework developed in this paper allows for other and more general loss functions. However, we argue that this relatively parsimonious loss function specification might provide an adequate approximation of the U.S. Administration's loss function. Firstly, it is likely that a politician has asymmetric preferences over budget deficit forecast errors as it is politically costly to cut government consumption or to raise taxes. Therefore, it might be relatively less costly to under-predict the budget surplus than to over-predict it. Secondly, the politician's preferences might differ at short and long forecast horizons as there are a frequent substitutions in politics. For example, a U.S. president in his second election period knows that he will not be in office after his second period runs out and this might be reflected in his preferences in the sense that is relatively less "costly" to overpredict the budget surplus at long forecast horizons.

1.4 Simulation Results

The purpose of this section is to briefly examine the finite sample behavior of the proposed estimator of $\boldsymbol{\alpha}_0$ and the forecast rationality test in a Monte Carlo experiment. We focus on a simple AR(1) model

$$Y_t = \mu + \theta Y_{t-1} + \varepsilon_t \ , \ \varepsilon_t|\mathcal{F}_{t-1} \sim i.i.d\mathcal{N}(0,\sigma^2) \ , \ t = 1,...,T$$

where Y_t is the economic variable of interest, $(\mu,\theta,\sigma^2)^T$ are parameters and ε_t is an unobserved error term. We construct samples of size $\tau = R + T + (H-1)$ for $\mu = 1$, $\theta = 0.90$ and $\sigma^2 = 1$. The forecaster uses a recursive window, with initial size $R = 100$, to construct T $H-period$ ahead forecasts, $\hat{\mathbf{Y}}^*_{t+H,t} = (\hat{Y}^*_{t+1,t},...,\hat{Y}^*_{t+H,t})^T$.

Under the given model specification, we have that

$$\hat{\epsilon}_{t+h,t}(\hat{Y}_{t+h,t})|\mathcal{F}_t \sim \mathcal{N}\left(\underbrace{\theta^h Y_t + (1-\theta^h)\frac{\mu}{1-\theta} - \hat{Y}_{t+h,t}}_{\mu_{\hat{\epsilon}_{t+h,t}}}, \underbrace{\sigma^2\left(\frac{1-\theta^{2h}}{1-\theta^2}\right)}_{\sigma^2_{\hat{\epsilon}_{t+h,t}}}\right) \ , \ h = 1,...,H$$

and this implies that $|\hat{\epsilon}_{t+h,t}|\,|\mathcal{F}_{t-1}$ has a folded normal distribution with mean given by $\sqrt{\sigma^2_{\hat{\epsilon}_h}\frac{2}{\pi}}exp(-\mu^2_{\hat{\epsilon}_{t+h,t}}/2\sigma^2_{\hat{\epsilon}_{t+h,t}}) + \mu_{\hat{\epsilon}_h}[1-2\Phi(-\mu_{\hat{\epsilon}_{t+h,t}}/\sigma_{\hat{\epsilon}_{t+h,t}})]$ where Φ denotes the cumulative distribution function of a standard normal random variable. Under mean squared error loss, the optimal forecast path is given by

$$\mathbf{Y}^*_{t+H,t} = \begin{bmatrix} \hat{\theta} \\ \hat{\theta}^2 \\ \vdots \\ \hat{\theta}^H \end{bmatrix} Y_t$$

There does not exist a closed form solution for the optimal term structure of point forecast under LHVQ preferences, but they are defined as the solution to the following set of equations,

$$E[\hat{\epsilon}_{t+h,t}(\hat{Y}_{t+h,t})|\mathcal{F}_t] = (1-f_h(\boldsymbol{\alpha}_0))(\sqrt{\hat{\sigma}^2_{\hat{\epsilon}_1}\frac{2}{\pi}}exp(-\hat{\mu}^2_{\hat{\epsilon}_{t+h,t}}/2\hat{\sigma}^2_{\hat{\epsilon}_1}) + \hat{\mu}_{\hat{\epsilon}_{t+h,t}}[1-2\Phi(-\hat{\mu}_{\hat{\epsilon}_{t+h,t}}/\hat{\sigma}_{\hat{\epsilon}_1})],$$

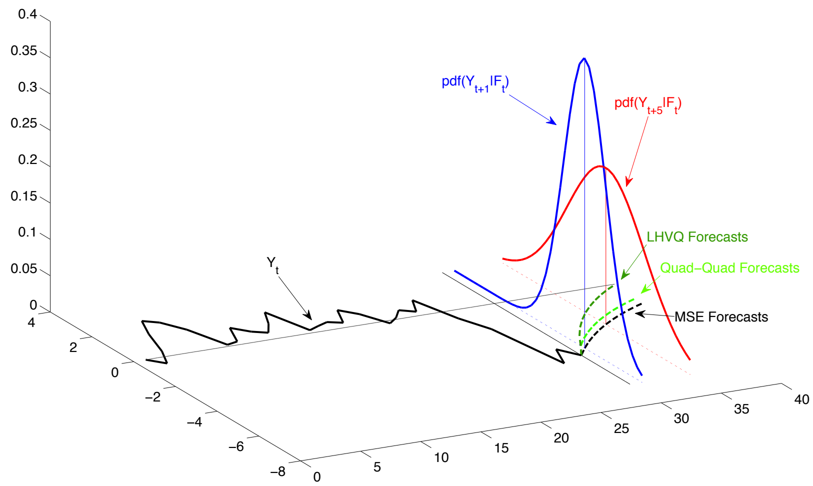


Figure 1.1: *Illustration of optimal forecast-path under MSE, Quad-Quad and LHVQ preferences*

for $h = 1, ..., H$. Figure 1.1 illustrates a simulation of the process and the optimal term structure of point forecasts under MSE, Quad-Quad ($\alpha = 60\%$) and LHVQ $((\alpha_{0,1}, \alpha_{0,2})^T = (60\%, 20\%)^T)$ preferences.

As we will discuss in more detail in section 1.5, it is clear that the bias term is an increasing function of the forecast horizon for *both* Quad-Quad and LHVQ preferences. However, the bias increases faster for the agent with LHVQ preferences which reflects that the asymmetry increases with the forecast horizon (as $\alpha_{0,2} > 0$).

In order to assess the finite sample properties of the rationality test, we simulate the process 1.000 times and construct 25, 50, 100 and 250 term structure forecasts ($H = 3$) for each simulation based on Quad-Quad ($\alpha_0 = 60\%$) and LHVQ preferences ($\alpha_{0,1} = 60\%$ and $\alpha_{0,2} = 20\%$). The results are shown in table 1.1. For the forecast formed by the agent with LHVQ preferences, the forecast rationality test based on a LHVQ specification appears to do a reasonable job and, in line with our expectations, forecast rationality can generally not be rejected. That said, when the number of forecasts is relatively low, the test tends to be oversized. In contrast, the forecast rationality test based on a Quad-Quad specification tends *a) Null Hypothesis: LHVQ preferences with* $(\alpha_{0,1}, \alpha_{0,2})^T = (60\%, 20\%)^T$.

	T=25	T=50	T=100	T=250
LHVQ Specification (Size)	15.1%	11.9%	7.3%	4.3%
Quad-Quad Specification ("Power")	42.4%	44.9%	77.9%	99.9%

b) Null Hypothesis: Quad-Quad preferences with $\alpha_0 = 60\%$.

	T=25	T=50	T=100	T=250
LHVQ Specification (Size)	24.1%	12.8%	8.2%	4.2%
Quad-Quad Specification (Size)	25.6%	15.7%	8.3%	3.7%

Table 1.1: *Results from simulation experiment.*

to falsely reject forecast rationality. For example, for $P = 25$, the test tends to reject forecast rationality for more than 40% of the simulations. This clearly illustrates that a multi-horizon forecast rationality test, based on a quad-quad preference specification, can lead to a false rejection of forecast rationality.

Based on the forecasts formed by the agent with Quad-Quad preferences, both test appears to do a reasonable job. However, both tests are again oversized when the number of forecasts is relatively low.

Finally, figure 1.2 illustrates the distribution of the proposed estimator of the loss function shape parameters, based on the LHVQ specification. Based on the simulation, it is clear that the estimator is consistent for T going towards infinity.

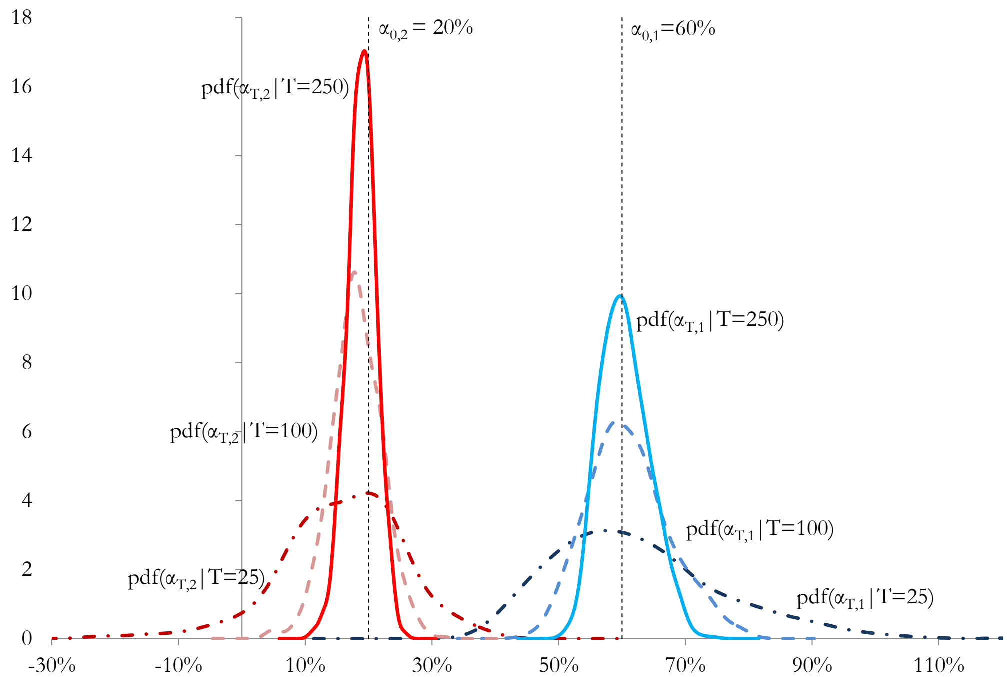


Figure 1.2: *Distribution of loss function estimator based on LHVQ specification.*

1.5 Inferring the U.S. Administration's loss function

One of the main events in the US political calendar is the annual publication of the federal budget for the upcoming fiscal year accompanied by projections of the federal budget deficit years ahead. While such budget projections are inevitably inaccurate the question remains whether they are also subject to political manipulation. In the theory section, we provided conditions for identification, estimation and inference of the shape parameters of an agent's loss function. In this section, we employ this methodology to estimate the U.S. Administration's loss function, based on it's term structure of budget surplus forecasts over the period 1975-2012, and test whether the forecasts are rational.[2]

As mentioned in section 1.2, the empirical analysis is motivated by the political economy literature which suggest that there are reasons to believe that a government's preferences over budget deficit forecast errors are not symmetric and, furthermore, that the relative costs

[2]To be precise, the forecasts are produced by the Office of Management and Budget, the largest office within the Executive Office of the President of the United States (EOP).

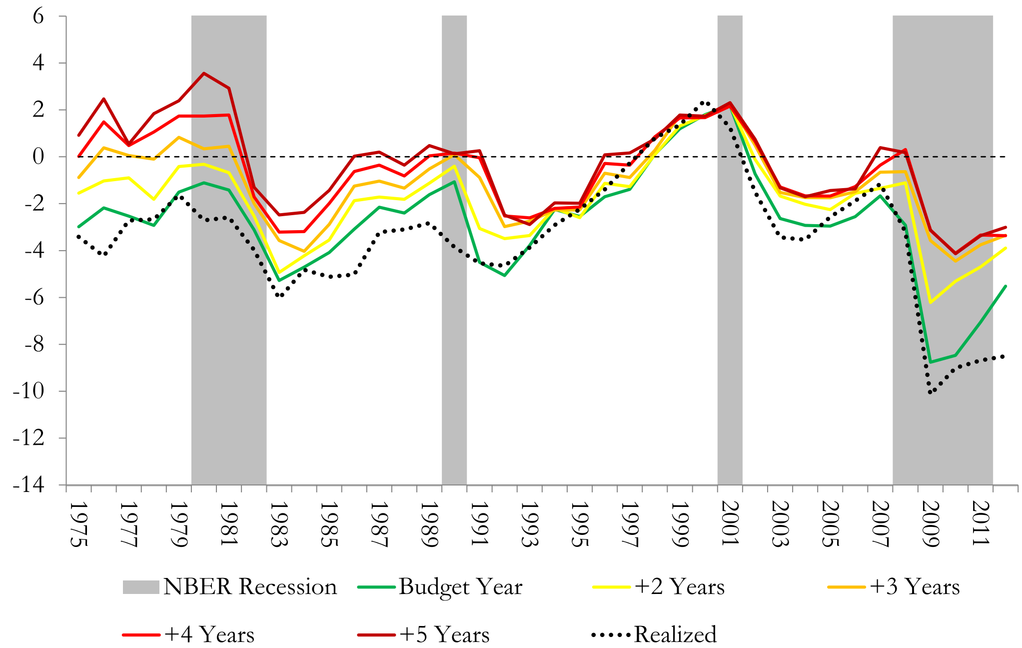


Figure 1.3: *U.S. Budget Surplus Forecasts, 1975-2012. The term structure of point forecasts are formed in January or February by the Office of Management and Budget (OMB) and the fiscal year begins the 1st of October. The dates of the forecasts refer to the year they was formed. NBER recessions are highlighted in gray.*

might depend on the forecast horizon. For example, one could argue that it is politically costly to cut government spending, or to raise taxes, and this implies that the relative costs of positive forecast errors, i.e. if the budget surplus is larger than expected, are higher than negative forecast errors. Furthermore, if the costs associated with observing a forecast error depends on whether a politician is in office or not, it is likely that the loss function depends on the forecast horizon.

Figure 1.3 illustrates the realized U.S. budget surplus and the U.S. Administration's term structure of point forecasts over the period 1975-2012. At a first glance, the U.S. budget surplus forecasts appear to be optimistic relative to the realized values denoted by the dotted black line. Also, interestingly, the term structure of the President's budget forecasts is upward sloping or flat during the entire period.

Based on this term structure of point forecasts, we estimate the U.S. Administration's preferences based on different loss function specifications. Firstly, we simply assume that the preferences can be approximated by MSE or MAE and test for rationality using an overidentified J-test. As illustrated in table 1.2, rationality is rejected at all conventional significance levels for both the MSE and MAE loss function specification.[3] This result might indicate that the U.S. administration does not form rational forecasts. However, it could also indicate that they have asymmetric preferences over budget deficit forecast errors. In order to evaluate the latter interpretation further we estimate the loss function shape parameters based on a Quad-Quad and a Lin-Lin loss function specification. We find that the preferences are asymmetric such that positive forecast errors are more costly than negative forecast errors. This could reflect that it is politically costly to cut government spending, or to raise taxes, and it might therefore be relative less "costly" for the U.S. Administration to underpredict the budget deficit. However, for both the Quad-Quad and the Lin-Lin loss function specification, we reject the joint hypothesis that forecasts are rational and that the loss function can be approximated by Quad-Quad or Lin-Lin preferences. If we believe that the U.S. Administration forms the budget surplus forecasts rationally, i.e. Assumption 1 holds, this suggests that both Quad-Quad or Lin-Lin are poor approximations of the true underlying loss function.

Even though both the Quad-Quad and Lin-Lin loss functions are relatively flexible, they do impose some structure on the loss function. For example, both Quad-Quad or Lin-Lin impose that the preferences are constant across forecast horizons. This might be a reasonable assumption in many applications but there are reasons to believe that it could be violated here. It is a well known fact that a politician rarely maintains the same position for a decade or longer. In particular, the U.S. president, who is heading the U.S. administration, is by law prohibited to be in office more than two election periods (e.g. around 8 years). Therefore, if the loss depends on whether the U.S. President is in office or not, the preferences at long forecast horizons might be different due to the higher probability that he/she is no longer in office. The loss functions LHVQ and LHVL, introduced in section 1.3, allow for this type of effect. Table 1.2 presents the estimation results when assuming that

[3] Mincer-Zarnowitz tests are also rejected for all forecast horizons. This is illustrated in Appendix C.

Loss Function	$\hat{\alpha}_{0,T}$	$\sqrt{Var[\hat{\alpha}_{0,T}]}$	$\hat{\alpha}_{1,T}$	$\sqrt{Var[\hat{\alpha}_{1,T}]}$	J test (p-val)
MSE	(50%)	-	-	-	<0.1%
MAE	(50%)	-	-	-	<0.1%
Quad-Quad	77.8%	8.60%	-	-	<0.1%
Lin-Lin	62.3%	9.69%	-	-	0.3%
LHVQ	60.0%	5.96%	16.0%	3.04%	49.84%
LHVL	55.1%	9.13%	7.6%	3.03%	91.3%

Table 1.2: *U.S. Government Loss Function Estimation Results. The GMM estimation is based on the following instruments* $\mathbf{V}_t = [1,\, Y_t]^T$*, 1975-2012.*

the U.S. Administration's preferences can be approximated by either LHVQ or LHVL. Interestingly, the empirical results indicate that the preferences appear to become more asymmetric at long forecast horizons and this effect is statistically significant for both model specifications. In addition, we cannot reject the joint test of forecast rationality and LHVQ or LHVL preferences. This suggests that the LHVQ and LHVL both provide a good approximation of the U.S. Administration's loss function. Figure 1.4 and 1.5 illustrate the estimated US Government loss function. The estimated loss function is also illustrated at different angles in Appendix C.

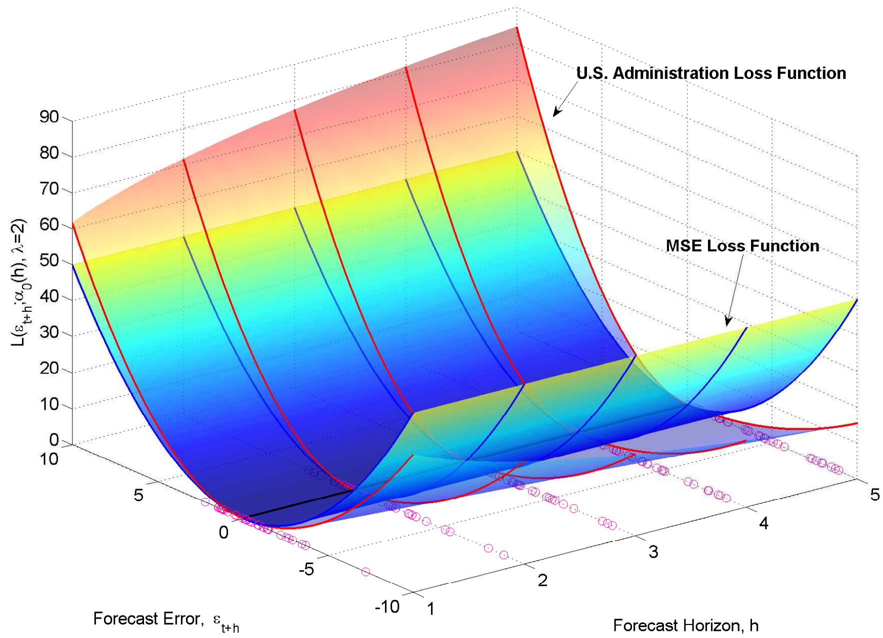


Figure 1.4: *Three dimensional illustration of the estimated loss function for the U.S. Administration. The GMM estimation is based on the following instruments* $\mathbf{V}_t = [1,\, Y_t]^T$ *, 1975-2012.*

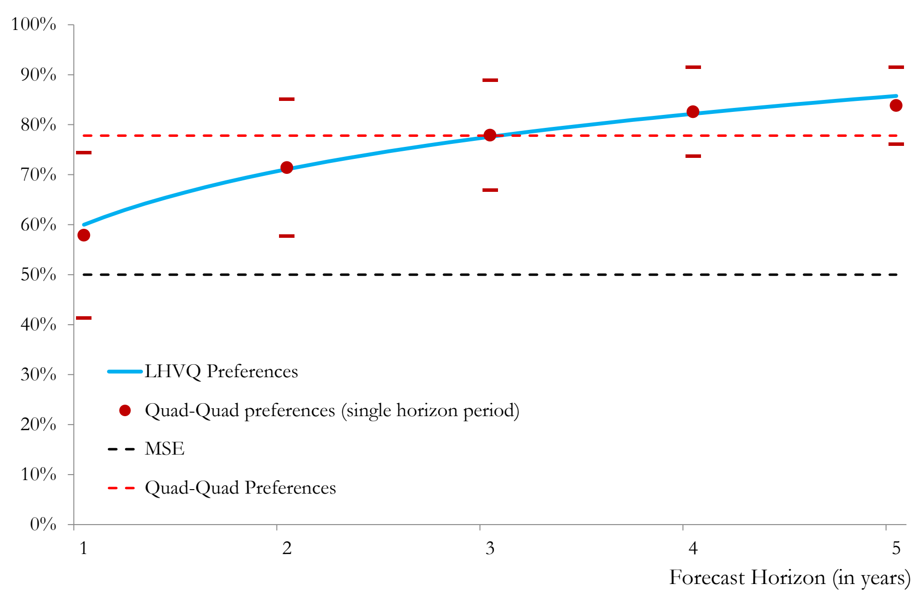


Figure 1.5: *Estimated loss function for the U.S. Administration based on Quad-Quad and LHVQ specification. The exhibit also illustrate the single period Quad-Quad loss function estimates for h=1,...,5 (red dots). The estimation is based on the following instruments* $\mathbf{V}_t = [1,\, Y_t]^T$ *, 1975-2012.*

Loss Function	$\hat{\alpha}_{0,T}$	$\sqrt{Var[\hat{\alpha}_{0,T}]}$	$\hat{\alpha}_{1,T}$	$\sqrt{Var[\hat{\alpha}_{1,T}]}$	J test (p-val)
First Election Period:					
LHVQ	58.7%	7.92%	10.6%	4.6%	23.3%
Second Election Period:					
LHVQ	53.9%	7.21%	32.8%	4.3%	39.2%

Table 1.3: *First and Second Election Period Estimation Results. The GMM estimation is based on the following instruments* $\mathbf{V}_t = [1,\, Y_t]^T$*, 1975-2012.*

1.5.1 The Impact of Election Uncertainty

"This is my last election. After my [second] election I have more flexibility." (U.S. President Barack Obama in a conversation with Russian leader Dmitry Medvedev, 26 March, 2012). The empirical results in the previous section suggest that the U.S. Administration's loss function is asymmetric and depends on the forecast horizon. We argued that the forecast horizon dependence could arise as a result of election uncertainty. In the United States, the president, who is heading the U.S. Administration, is prohibited by law to be in office more than two election periods. Therefore, if the forecast horizon dependence occurs because of election uncertainty, the loss function during the first and second election period should differ. In order to investigate this hypothesis further, we estimate the U.S. Administration's loss function based on two subsamples: one subsample where the current president is in his first election period and one subsample where the current president is in his second election period. The results are presented in table 1.3.

The results in table 1.3 provide two important insights. Firstly, the baseline preferences are almost identical for first and second election periods. If anything, presidents in their second term are slightly closer to symmetric loss in the short term forecasts. Since the positive forecast bias may be used as a device to increase popularity among the voters it is perhaps surprising that the baseline skewness of the second period loss function so closely resembles that of the first period loss function. Although the political party of the president may remain in office the president himself cannot be re-elected when the second term ends and we would therefore expect the gains from optimistic budget forecasts to be lower for second period presidents. However, an optimistic forecast may also be used as a tool to increase the flexibility of passing costly policies, and there is some evidence to suggest that second term politicians are prone to increase spending, eg. Besley & Case (1995). Studying U.S. Governors from 1950 to 1986 they conclude that "the main effect of term limits is to generate a fiscal cycle, with incumbents holding spending below the state's mean in their first term in office and spending significantly above the state's mean in the lame-duck term". If the same holds at the federal level, then it is not clear that the second period loss function should be close to symmetry.

Secondly, the forecast dependence is stronger based on the second election period. Positive forecast errors potentially give rise to two costs. Partly a loss of reputation when the optimism of the forecast is revealed, partly the need to cut spending to reduce the budget deficit resulting from lax policies that were previously justified by optimistic projections. A president taking office for a second time knows with certainty that he will no longer be in office when the current term ends. This means that the costs of forecasts extending beyond the second term of office will be substantially lower. The reputational cost will matter less and the task of cutting spending will fall on the successor. This should increase the asymmetry of the loss function over the forecast horizon which is supported by the results in table 1.3. Furthermore, policy takes time to implement. If a re-elected president in general pursues less budget friendly policies they may take effect only after some years. The forecast bias would then manifest itself with a time delay which would also amplify the positive bias over the forecast horizon.

1.5.2 Are the Results Driven by Large Negative Shocks?

Even though the estimation is based on a relative long sample from 1975-2012 there is still a danger that the results are driven by a few large shocks in the estimation period. In order to

Loss Function	$\hat{\alpha}_{0,T}$	$\sqrt{Var[\hat{\alpha}_{0,T}]}$	$\hat{\alpha}_{1,T}$	$\sqrt{Var[\hat{\alpha}_{1,T}]}$	J test (p-val)
MSE	(50%)	-	-	-	32.30%
MAE	(50%)	-	-	-	4.62%
Quad-Quad	46.03%	9.98%	-	-	97.71%
Lin-Lin	36.56%	6.66%	-	-	76.41%
LHVQ	42.42%	10.43%	12.37%	12.95%	97.17%
LHVL	31.91%	6.99%	14.04%	11.25%	98.40%

Table 1.4: *IMF Loss Function Estimation Results. The GMM estimation is based on the following instruments* $\mathbf{V}_t = [1,\, Y_t]^T$*, 1975-2012 (H=2).*

assess this potential issue further, we analyse how other institutions forecasted the U.S. budget surplus during the same period. If the results are driven by large unforecastable shocks, other institutions preferences will look similar. We choose to analyse how the IMF forecasted the U.S. budget deficit during the same period. IMF is an independent international organisation and is arguably under less political pressure relative to the U.S. Administration. The estimation results are presented in table 1.4.

Interestingly, we cannot reject that the IMF has MSE preferences over U.S. budget deficit errors. If anything, the empirical evidence suggests that the IMF are a bit conservative, i.e. weights negative forecast errors more heavily, based on the Quad-Quad or Lin-Lin loss function specification. It seems reasonable that the US Administration and the IMF would have roughly the same chances of predicting economic activity, or that the US Administration would have an informational advantage over the IMF. If the apparent optimistic bias in the US Administration's budget deficit forecasts was due to extraordinary occurences of large shocks suddenly driving the actual budget deficit below what was anticipated we would therefore also expect the same bias to be present in budget deficit forecasts made by the IMF. Since it appears that the IMF has close to symmetric preferences, as one would expect, this is suggestive evidence that the US Administration's optimistic bias is not due to an inability to capture negative shocks in the forecasts.

1.5.3 Bias Structure Interpretation under Quad-Quad Preferences

The empirical analysis in this paper has revealed that the U.S. Administration's preferences are asymmetric and depend on the forecast horizon. More specifically, we have found that positive forecast errors appear to be relatively more costly and that this asymmetry is stronger at longer forecast horizons. Intuitively, asymmetric preferences will imply that the expected forecast error, for an optimal forecast, is different from zero. Furthermore, intuition suggests that the expected forecast error bias is larger at higher forecast horizons if the preference asymmetry increases with the forecast horizons. However, while this intuitive interpretation is correct, there is also an additional reason why the expected forecast error bias is increasing with the forecast horizon, for $\lambda > 1$. In order to illustrate this it is useful to rewrite the unconditional moment conditions from Proposition 1 by using that $1\{Y_{t+h} - Y^*_{t+h,t} < 0\} = \frac{1}{2}\left(1 - [Y_{t+h} - Y^*_{t+h,t}]/|Y_{t+h} - Y^*_{t+h,t}|\right)$:

$$\begin{aligned}\mathbb{E}[diag(|\mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t}|^{\lambda-1})([\mathbf{Y}_{t+H} - \hat{\mathbf{Y}}^*_{t+H,t}]/|\mathbf{Y}_{t+H} - \hat{\mathbf{Y}}^*_{t+H,t}|)] = \\ \mathbb{E}[diag(|\mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t}|^{\lambda-1})(\boldsymbol{\iota}_H - 2\mathbf{f}(\boldsymbol{\alpha}, h))|\mathcal{F}_t]\end{aligned}$$

Under simple Quad-Quad preferences, $\mathbf{f}(\boldsymbol{\alpha}) = \alpha_0 \boldsymbol{\iota}_H$ and $\lambda = 2$, this translates into the following bias structure:

$$\begin{bmatrix} \mathbb{E}[\epsilon^*_{t+1,t}] \\ \mathbb{E}[\epsilon^*_{t+2,t}] \\ \vdots \\ \mathbb{E}[\epsilon^*_{t+H,t}] \end{bmatrix} = (1 - 2\alpha_0) \begin{bmatrix} \mathbb{E}[|\epsilon^*_{t+1,t}|] \\ \mathbb{E}[|\epsilon^*_{t+2,t}] \\ \vdots \\ \mathbb{E}[|\epsilon^*_{t+H,t}|] \end{bmatrix}$$

where $\epsilon^*_{t+h,t} = Y_{t+h} - Y^*_{t+h,t}$. Clearly, the expected forecast error depends on the asymmetry, $(1 - 2\alpha_0)$, *and* on the expected value of the absolute forecast errors. In order to interpret this result, it is useful to remember that the expected value of the absolute forecast errors is

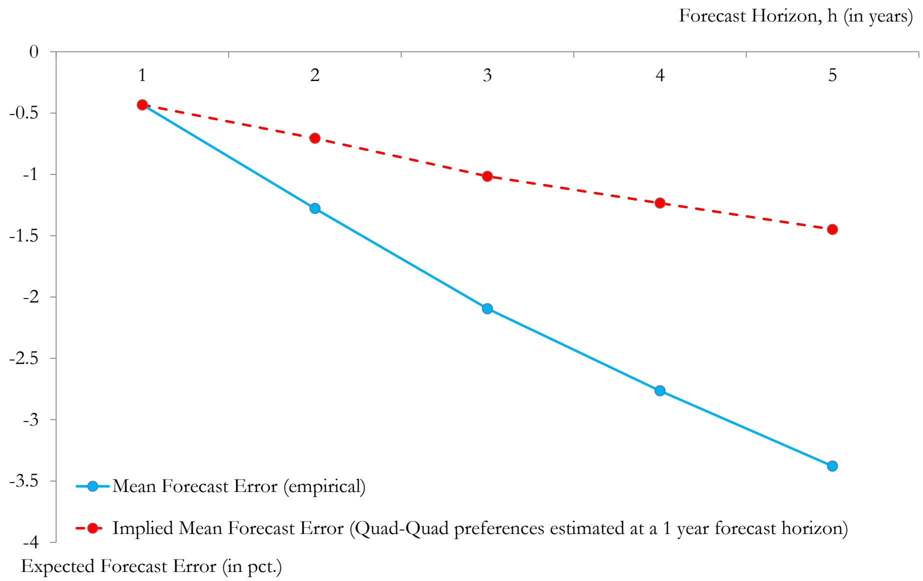


Figure 1.6: *Empirical and Quad-Quad implied bias structure based on U.S. Administration budget surplus forecasts, 1975-2012. The Quad-Quad implied bias structure is based on a single-horizon estimation for h=1 year.*

roughly related to the variance of the forecast error.[4] Therefore, as there is generally more uncertainty associated with forecasts at long forecast horizons, the variance of the forecast error is increasing with the forecast horizon. Hence, even under Quad-Quad preferences, the expected forecast error bias will be an increasing function of the forecast horizon since the forecast error variance increases. Figure 1.6 illustrates the implied bias structure under Quad-Quad preferences (estimated based on a one year forecast horizon) and compares it to the actual bias structure.

Figure 1.6 clearly illustrates that the expected forecast error bias, under Quad-Quad preferences, is an increasing function in the forecast horizon but that the realized forecast error bias is substantially bigger. In other words, this figure illustrates why the joint test of forecast rationality and Quad-Quad loss function preferences are rejected in table 1.2.

Frankel (2011), Frankel & Schreger (2012) and Kliesen & Thornton (2012) all note that the bias in budget forecasts display a tendency to increase with the forecast horizon. The

[4]For a Gaussian random variable, one can derive the following exact relationship: If $x \sim N(\mu, \sigma^2)$ then $E[|x|] = \sigma\sqrt{\frac{2}{\pi}} exp\{\frac{-\mu}{2\sigma^2}\} + \mu[1 - 2\Phi(\frac{-\mu}{\sigma})]$.

analysis above suggests that this pattern could arise under Quad-Quad preferences. One advantage of the methodology developed in this paper, is that it allows us to identify whether the increasing bias structure is simply a result of higher forecast error variance at long forecast horizons or whether it is due to more fundamental forecast-horizon varying features of the loss function. Based on the empirical analysis in this paper, features of the U.S. Administration's loss function appear to depend on the forecast horizon.

1.6 Conclusion

This paper provided conditions for identification and estimation of an agent's loss function based on an observed term structure of point forecasts. The main innovation of this approach is that we allow the loss function to belong to a flexible family of functions. In particular, this family of functions allows the preferences to be both asymmetric and to depend on the forecast horizon. In addition, we introduce a novel forecast rationality test which tests the joint hypothesis of forecast rationality and that the forecast error preferences can be approximated by this flexible class of loss functions. Based on a simulation experiment, we showed that the test has good size properties under the null where a rational agent's preferences depend on the forecast horizon. We employ the approach to analyse the U.S. Government's preferences over budget surplus forecast errors. Interestingly, we find that it is relatively more costly for the government to underestimate the budget surplus and that this asymmetry is stronger at long forecast horizons. We conjectured that this may reflect a political bias that arises because it is politically costly to cut government expenses, or to raise taxes, and that this effect is stronger at long forecast horizons due to uncertainty associated with political elections.

1.7 Appendices

Appendix A: Assumptions

Notation:

The conditional distribution of Y_{t+h} given $\mathcal{F}_t$ is denoted by $F_{t+h,t}$ and the conditional density is denoted by $f_{t+h,t}$. Given $\lambda \in \mathbb{R}^+$and $\mathbf{f}\colon \Theta \to (0,1)^H$ we let $g(\boldsymbol{\epsilon}^*_{t+H,t}; \boldsymbol{\alpha}) = diag(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1})(1\{\boldsymbol{\epsilon}^*_{t+H,t} < 0\} - \mathbf{f}(\boldsymbol{\alpha}))$, $\boldsymbol{\epsilon}^*_{t+H,t} = \mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t}$, $\mathbf{Y}^*_{t+H,t} = \underset{\hat{\mathbf{Y}}_{t+H,t} \in \mathcal{Y}}{arg\,min}\, \mathbb{E}[\mathcal{L}(\hat{\boldsymbol{\epsilon}}_{t+H,t}; \boldsymbol{\alpha}_0, \lambda_0)|\mathcal{F}_t]$ and $\mathcal{L}(\boldsymbol{\epsilon}^*_{t+H,t}; \boldsymbol{\alpha}_0; \lambda_0) = \mathcal{L}_{t+H}$. Furthermore, we let $h_{t+H}(\boldsymbol{\alpha}) = \mathbf{Z}_t \otimes diag(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1})(1\{\boldsymbol{\epsilon}^*_{t+H,t} < 0\} - \mathbf{f}(\boldsymbol{\alpha}))$ and $\hat{h}_{t+H}(\boldsymbol{\alpha}) = \mathbf{Z}_t \otimes diag(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1})(1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t} < 0\} - \mathbf{f}(\boldsymbol{\alpha}))$ where $\hat{\boldsymbol{\epsilon}}^*_{t+H,t} = \mathbf{Y}_{t+H} - \hat{\mathbf{Y}}^*_{t+H,t}$ and $\hat{\mathbf{Y}}^*_{t+H,t} = \underset{\hat{\mathbf{Y}}_{t+H,t} \in \mathcal{Y}}{arg\,min}\, \hat{\mathbb{E}}[\mathcal{L}(\hat{\boldsymbol{\epsilon}}_{t+H,t}; \boldsymbol{\alpha}_0, \lambda_0)|\mathcal{F}_t]$. . Finally, for $\mathbf{x} = (x_1, ..., x_N)^T \in \mathbb{R}^N$, we use $diag(\mathbf{x})$ to define the corresponding diagonal matrix,

$$diag(\mathbf{x}) = \begin{pmatrix} x_1 & 0 & \dots & 0 \\ 0 & x_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \dots & & x_N \end{pmatrix}.$$

Assumption 3 (Optimality and Identification):

A. *$\mathcal{Y}$ is a compact set in $\mathbb{R}^H$ and $\mathbf{Y}^*_{t+H,t}$ is interior to $\mathcal{Y}$, i.e. $\mathbf{Y}^*_{t+H,t} \in \mathring{\mathcal{Y}}$.*

B. *The conditional density of Y_{t+h} is strictly positive for all h, i.e. for every $x \in \mathbb{R}$, $f_{t+h,t}(x) > 0\, \forall h \in \{1, ..., H\}$.*

C. *The $m \times 1$ vector $\mathbf{V}_t$ (with first component 1) is such that, given $\lambda \in \mathbb{R}^+$ and $\forall \alpha_0 \in \mathring{\Theta}$,*

$$\mathbb{E}[\mathbf{V}_t \otimes \left|\mathbf{Y}_{t+h} - \mathbf{Y}^*_{t+H,t}\right|^{\lambda-1}] \neq 0 \; element\, by\, element$$

D. The process $\{(\boldsymbol{\epsilon}^{*T}_{t+H}, \mathbf{z}_t)\}$ is strictly stationary.

Assumption 4 (Consistency):

A. The process $\{(\hat{\boldsymbol{\epsilon}}^{*T}_{t+H}, \mathbf{z}_t)\}$ is strictly stationary and $\alpha - mixing$ with mixing coefficient α of size $-r/(r-1)$, $r > 1$, and given λ, $1 \leq \lambda < \infty$, there exist some $\varepsilon > 0$, $\Delta_1 > 0$ and $\Delta_2 > 0$ such that

$$\mathbb{E}[|\hat{\boldsymbol{\epsilon}}^{*}_{t+H,t}|^{(\lambda-1)(2r+\varepsilon)}) \leq \Delta_1 < \infty$$
$$\mathbb{E}[|\mathbf{Z}_t|^{(2r+\varepsilon)})] \leq \Delta_2 < \infty$$

B. The weight matrix $\mathbf{W}_T$ is positive semidefinite and converges in probability to a positive definite matrix, $\mathbf{W}$.

C. $\{\hat{Q}_T(\boldsymbol{\alpha}) : T \geq 1\}$is stochastically equicontinuous on Θ.

D. For any $\varepsilon > 0$, and every $R \leq t \leq T$, that

$$lim_{T\rightarrow\infty} Pr(|\hat{\boldsymbol{\epsilon}}^{*}_{t+H,t} - \boldsymbol{\epsilon}^{*}_{t+H,t}| > \varepsilon) = 0$$

where $R \rightarrow \infty$ as $T \rightarrow \infty$.

Assumption 5 (Asymptotic Normality):

A. The gradient of $\mathbb{E}[h_{t+H}(\boldsymbol{\alpha})] = \mathbb{E}[\mathbf{Z}_t \otimes diag(|\boldsymbol{\epsilon}^{*}_{t+H,t}|^{\lambda-1})(1\{\boldsymbol{\epsilon}^{*}_{t+H,t} < 0\} - \mathbf{f}(\boldsymbol{\alpha}))]$ is continuous on Θ, it exist and it is finite,

$$\begin{aligned} \frac{\partial \mathbb{E}[h_{t+h}(\boldsymbol{\alpha})]}{\partial \boldsymbol{\alpha}} &= -\mathbb{E}\left[\mathbf{Z}_t \otimes diag(|\boldsymbol{\epsilon}^{*}_{t+H,t}|^{\lambda-1})\frac{\partial f(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right] < \infty \\ &= -\mathbf{m}(\boldsymbol{\alpha}). \end{aligned}$$

In addition, we assume that a law of large number applies such that

$$\frac{1}{T}\sum_{t=1+R}^{R+T} \mathbf{Z}_t \otimes diag(|\hat{\boldsymbol{\epsilon}}^{*}_{t+H,t}|^{\lambda-1})\frac{\partial f(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \quad \overset{p}{\rightarrow} \quad \mathbb{E}\left[\mathbf{Z}_t \otimes diag(|\boldsymbol{\epsilon}^{*}_{t+H,t}|^{\lambda-1})\frac{\partial f(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right] \; as \; T \rightarrow \infty$$

where it is assumed that $R \rightarrow \infty$ as $T \rightarrow \infty$.

B. $lim_{T\rightarrow\infty} Var[\sqrt{T}\mathbf{h}(\boldsymbol{\alpha}_0)] = \mathbf{S}$ exists and is a finite valued positive definite matrix. Furthermore, $\mathbf{W}_T$ is a consistent estimator of $\mathbf{S}^{-1}$.

C. The process $\{(\hat{\boldsymbol{\epsilon}}^{*T}_{t+H}, \mathbf{z}_t)\}$ is strictly stationary and $\alpha - mixing$ with mixing coefficient α of size $-r/(r-2)$, $r > 2$. Furthermore, given λ, $1 \leq \lambda < \infty$, there exist $\Delta_3 > 0$ such that

$$\mathbb{E}\left[\left|\mathbf{Z}_t \otimes diag\left(|\hat{\boldsymbol{\epsilon}}^{*}_{t+H,t}|^{\lambda-1}\right)(1\{\hat{\boldsymbol{\epsilon}}^{*}_{t+H,t} < 0\} - f(\boldsymbol{\alpha}, h))\right|^{r}\right] \leq \Delta_3 < \infty$$

for some $r \geq 2$.

Appendix B: Proofs

Proof of Proposition 1: We first show that (1.3.2) is a necessary condition for optimality of $\mathbf{Y}^*_{t+H,t}$. From assumption 1 and 2 we know that the optimal forecast is a solution to:

$$\mathbf{Y}^*_{t+H,t} = \underset{\hat{\mathbf{Y}}_{t+H,t} \in \mathcal{Y}}{arg\,min}\, \mathbb{E}[\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t}, \mathbf{Y}_{t+H,t}; \boldsymbol{\alpha}_0)|\mathcal{F}_t]$$

where $\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t}, \mathbf{Y}_{t+H,t}; \boldsymbol{\alpha}_0) = \sum_{h=1}^{H}[f(\boldsymbol{\alpha}, h) + (1 - 2f(\boldsymbol{\alpha}, h)) \cdot 1\{\hat{\epsilon}_{t+h,t} < 0\}]|\hat{\epsilon}_{t+h,t}|^{\lambda}$. First, we show that $\mathbb{E}[\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t}, \mathbf{Y}_{t+H,t}; \boldsymbol{\alpha}_0)|\mathcal{F}_t]$ is a continuously differentiable function on $\mathcal{Y}$. $\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t}, \mathbf{Y}_{t+H,t}; \boldsymbol{\alpha}_0)$ is a differentiable function on $\mathcal{Y}\backslash A_{t+H}$, where $A_{t+H} = \{\hat{\mathbf{Y}}_{t+H,t} \in \mathcal{Y} : \hat{Y}_{t+1,t} = Y_{t+1,t} \cup \hat{Y}_{t+2,t} = Y_{t+2,t} \cup ... \cup \hat{Y}_{t+H,t} = Y_{t+H,t}\}$. By the law of iterated expectations, $\mathbb{E}[\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t}, \mathbf{Y}_{t+H,t}; \boldsymbol{\alpha}_0)|\mathcal{F}_t] = \mathbb{E}\left[\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t}, \mathbf{Y}_{t+H,t}; \boldsymbol{\alpha}_0)\mathbb{E}[1|\mathcal{F}_{t+H}]|\mathcal{F}_t\right]$ as $\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t}, \mathbf{Y}_{t+H,t}; \boldsymbol{\alpha}_0) \in \mathcal{F}_{t+H}$. Furthermore, $\mathbb{E}[1|\mathcal{F}_{t+H}] = \mathbb{E}[\mathbf{1}\{\hat{\mathbf{Y}}_{t+H,t} \in A_{t+H}\}|\mathcal{F}_{t+H}] + \mathbb{E}[\mathbf{1}\{\hat{\mathbf{Y}}_{t+H,t} \in A^c_{t+H}\}|\mathcal{F}_{t+H}] = \mathbb{E}[\mathbf{1}\{\hat{\mathbf{Y}}_{t+H,t} \in A^c_{t+H}\}|\mathcal{F}_{t+H}]$ as $\mathbb{E}[\mathbf{1}\{\hat{\mathbf{Y}}_{t+H,t} \in A_{t+H}\}|\mathcal{F}_{t+H}] = 0$. Hence, $\mathbb{E}[\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t}, \mathbf{Y}_{t+H,t}; \boldsymbol{\alpha}_0)\mathcal{F}_t] = \mathbb{E}\left[\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t})\mathbb{E}[\mathbf{1}\{\hat{\mathbf{Y}}_{t+H,t} \in X^c_{t+H}\}|\mathcal{F}_{t+H}]|\mathcal{F}_t\right]$ is a continuously differentiable function on $\mathcal{Y}$.

Now we analyse the gradient of the objective function:

$$\mathbb{E}[\nabla_{\hat{\mathbf{Y}}}\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t}, \mathbf{Y}_{t+H,t}; \boldsymbol{\alpha}_0)|\mathcal{F}_t] = \begin{bmatrix} -\lambda\mathbb{E}\left[(f(\boldsymbol{\alpha},1) + (1-2f(\boldsymbol{\alpha},1)) \cdot 1\{Y_{t+1} - \hat{Y}_{t+1,t} < 0\})]|Y_{t+1} - \hat{Y}_{t+1,t}|^{\lambda-1}\left(\frac{Y_{t+1}-\hat{Y}_{t+1,t}}{|Y_{t+1}-\hat{Y}_{t+1,t}|}\right)|\mathcal{F}_t\right] \\ +(1-2f(\boldsymbol{\alpha},1))\mathbb{E}\left[\frac{\partial 1\{Y_{t+1}-\hat{Y}_{t+1,t}<0\}}{\partial \hat{Y}_{t+1,t}}]|Y_{t+1} - \hat{Y}_{t+1,t}|^{\lambda}|\mathcal{F}_t\right] \\ -\lambda\mathbb{E}[\left[(f(\boldsymbol{\alpha},2) + (1-2f(\boldsymbol{\alpha},2)) \cdot 1\{Y_{t+2} - \hat{Y}_{t+2,t} < 0\})]|Y_{t+2} - \hat{Y}_{t+2,t}|^{\lambda-1}\left(\frac{Y_{t+2}-\hat{Y}_{t+2,t}}{|Y_{t+2}-\hat{Y}_{t+2,t}|}\right)|\mathcal{F}_t\right] \\ +(1-2f(\boldsymbol{\alpha},2))\mathbb{E}\left[\frac{\partial 1\{Y_{t+2}-\hat{Y}_{t+2,t}<0\}}{\partial \hat{Y}_{t+2,t}}]|Y_{t+2} - \hat{Y}_{t+2,t}|^{\lambda}|\mathcal{F}_t\right] \\ \vdots \\ -\lambda\mathbb{E}\left[(f(\boldsymbol{\alpha},H) + (1-2f(\boldsymbol{\alpha},H)) \cdot 1\{Y_{t+H} - \hat{Y}_{t+H,t} < 0\})]|Y_{t+H} - \hat{Y}_{t+H,t}|^{\lambda-1}\left(\frac{Y_{t+H}-\hat{Y}_{t+H,t}}{|Y_{t+H}-\hat{Y}_{t+H,t}}\right)|\mathcal{F}_t\right] \\ +(1-2f(\boldsymbol{\alpha},H))\mathbb{E}\left[\frac{\partial 1\{Y_{t+H}-\hat{Y}_{t+H,t}<0\}}{\partial \hat{Y}_{t+H,t}}]|Y_{t+H} - \hat{Y}_{t+H,t}|^{\lambda}|\mathcal{F}_t\right] \end{bmatrix}$$

Note that $\frac{\partial 1\{Y_{t+h}-\hat{Y}_{t+h,t}<0\}}{\partial \hat{Y}_{t+h,t}} = \delta(\hat{Y}_{t+h,t} - Y_{t+h})$, where $\delta(x) = \underset{a\to\infty}{lim}\left(\frac{1}{a\sqrt{\pi}}e^{-x^2/a^2}\right)$ is the Dirac function. Therefore, $\mathbb{E}\left[\delta(\hat{Y}_{t+h,t} - Y_{t+h})|Y_{t+h} - \hat{Y}_{t+h,t}|^{\lambda}\right] = 0$ for any non-zero λ.
Furthermore, $\left(\frac{Y_{t+h}-\hat{Y}_{t+h,t}}{|Y_{t+h}-\hat{Y}_{t+h,t}|}\right) = \left(1 - 2\cdot \mathbf{1}\{Y_{t+h} - \hat{Y}_{t+h,t} < 0\}\right)$ such that
$\left(\frac{Y_{t+h}-\hat{Y}_{t+h,t}}{|Y_{t+h}-\hat{Y}_{t+h,t}|}\right)(f(\boldsymbol{\alpha},h)+(1-2f(\boldsymbol{\alpha},h)\mathbf{1}\{Y_{t+h}-\hat{Y}_{t+h,t}<0\}) = (f(\boldsymbol{\alpha},h)-1\{Y_{t+h}-\hat{Y}_{t+h,t}<0\})$.
Thus,

$$\begin{aligned}\mathbb{E}[\nabla_{\hat{\mathbf{Y}}}\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t},\mathbf{Y}_{t+H,t};\boldsymbol{\alpha}_0)|\mathcal{F}_t] &= \begin{bmatrix}\lambda\mathbb{E}[(1\{Y_{t+1}-\hat{Y}_{t+1,t}<0\}-f_1(\boldsymbol{\alpha}_0)]|Y_{t+1}-\hat{Y}_{t+1,t}|^{\lambda-1}|\mathcal{F}_t]\\ \lambda\mathbb{E}[(1\{Y_{t+2}-\hat{Y}_{t+2,t}<0\}-f_2(\boldsymbol{\alpha}_0)]|Y_{t+2}-\hat{Y}_{t+2,t}|^{\lambda-1}|\mathcal{F}_t]\\ \vdots\\ \lambda\mathbb{E}[(1\{Y_{t+H}-\hat{Y}_{t+H,t}<0\}-f_H(\boldsymbol{\alpha}_0)]|Y_{t+H}-\hat{Y}_{t+H,t}|^{\lambda-1}|\mathcal{F}_t]\end{bmatrix}\\ &= \lambda\mathbb{E}[diag(|\mathbf{Y}_{t+H}-\hat{\mathbf{Y}}_{t+H,t}|^{\lambda-1})(1\{\mathbf{Y}_{t+H}-\hat{\mathbf{Y}}_{t+H,t}<0\}-\mathbf{f}(\boldsymbol{\alpha}_0))|\mathcal{F}_t]\end{aligned}$$

By A.3 we know that $\mathbf{Y}^*_{t+H,t} \in \overset{\circ}{\mathcal{Y}}$. Therefore, if $\mathbf{Y}^*_{t+H,t}$ is the minimum of $\mathbb{E}[\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t},\mathbf{Y}_{t+H,t};\boldsymbol{\alpha}_0)|\mathcal{F}_t]$, then $\mathbf{Y}^*_{t+H,t}$ is a solution to $\mathbb{E}[\nabla_{\hat{\mathbf{Y}}}\mathcal{L}(\mathbf{Y}^*_{t+H,t},\mathbf{Y}_{t+H,t};\boldsymbol{\alpha}_0)|\mathcal{F}_t] = 0$ (e.g. by theorem 3.7.13 in Schwartz (1997, vol. 2, p. 168)). This completes the necessary part of the proof.

We now derive a set of sufficient conditions for $\mathbf{Y}^*_{t+H,t} \in \overset{\circ}{\mathcal{Y}}$ to be a solution to the minimization problem in (1.3.2). We will use the results that $\hat{\mathbf{Y}}^*_{t+H,t}$ is a strict minimum of $\mathbb{E}[\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t},\mathbf{Y}_{t+H,t};\boldsymbol{\alpha}_0)|\mathcal{F}_t]$ on $\overset{\circ}{\mathcal{Y}}$ if $\mathbb{E}[\nabla_{\hat{\mathbf{Y}}}\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t},\mathbf{Y}_{t+H,t};\boldsymbol{\alpha}_0)|\mathcal{F}_t] = 0$ and $\mathbb{E}[\nabla_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t},\mathbf{Y}_{t+H,t};\boldsymbol{\alpha}_0)|\mathcal{F}_t]$ is positive definite (e.g. theorem 3.7.13 in Schwartz (1997, vol. 2, p 169)). In order to show that $\mathbb{E}[\nabla_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t},\mathbf{Y}_{t+H,t};\boldsymbol{\alpha}_0)|\mathcal{F}_t]$ is positive definite, we will use the well known result that a matrix is positive definite if and only if all its eigenvalues are strictly positive. $\mathbb{E}[\nabla_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t})|\mathcal{F}_t]$ can be written as

$$\mathbb{E}[\nabla_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t},\mathbf{Y}_{t+H,t};\boldsymbol{\alpha}_0)|\mathcal{F}_t] = \lambda(\lambda-1)\mathbb{E}[diag(\kappa)|\mathcal{F}_t] + \lambda\mathbb{E}[diag(\nu)|\mathcal{F}_t],$$

where

$$\kappa = \begin{bmatrix} (1\{Y_{t+1}-\hat{Y}_{t+1,t}<0\}-f_1(\boldsymbol{\alpha}_0))\left(\frac{Y_{t+1}-\hat{Y}_{t+1,t}}{|Y_{t+1}-\hat{Y}_{t+1,t}|}\right)|Y_{t+1}-\hat{Y}_{t+1,t}|^{\lambda-2} \\ (1\{Y_{t+2}-\hat{Y}_{t+2,t}<0\}-f_2(\boldsymbol{\alpha}_0))\left(\frac{Y_{t+2}-\hat{Y}_{t+2,t}}{|Y_{t+2}-\hat{Y}_{t+2,t}|}\right)|Y_{t+2}-\hat{Y}_{t+2,t}|^{\lambda-2} \\ \vdots \\ (1\{Y_{t+H}-\hat{Y}_{t+H,t}<0\}-f_H(\boldsymbol{\alpha}_0))\left(\frac{Y_{t+H}-\hat{Y}_{t+H,t}}{|Y_{t+H}-\hat{Y}_{t+H,t}|}\right)|Y_{t+H}-\hat{Y}_{t+H,t}|^{\lambda-2} \end{bmatrix}$$

$$= \begin{bmatrix} [f_1(\boldsymbol{\alpha}_0)+(1-2f_1(\boldsymbol{\alpha}_0))1\{Y_{t+1}-\hat{Y}_{t+1,t}<0\}]|Y_{t+1}-\hat{Y}_{t+1,t}|^{\lambda-2} \\ [f_2(\boldsymbol{\alpha}_0)+(1-2f_2(\boldsymbol{\alpha}_0))1\{Y_{t+2}-\hat{Y}_{t+2,t}<0\}]|Y_{t+2}-\hat{Y}_{t+2,t}|^{\lambda-2} \\ \vdots \\ [f_H(\boldsymbol{\alpha}_0)+(1-2f_H(\boldsymbol{\alpha}_0))1\{Y_{t+H}-\hat{Y}_{t+H,t}<0\}]|Y_{t+H}-\hat{Y}_{t+H,t}|^{\lambda-2} \end{bmatrix}$$

and

$$\nu = \begin{bmatrix} \delta(\hat{Y}_{t+1,t}-Y_{t+1})|Y_{t+1}-\hat{Y}_{t+1,t}|^{\lambda-1} \\ \delta(\hat{Y}_{t+2,t}-Y_{t+2})|Y_{t+2}-\hat{Y}_{t+2,t}|^{\lambda-1} \\ \vdots \\ \delta(\hat{Y}_{t+H,t}-Y_{t+H})|Y_{t+H}-\hat{Y}_{t+H,t}|^{\lambda-1} \end{bmatrix}$$

Consider the following two cases separately: $\lambda = 1$ and $\lambda > 1$. For $\lambda = 1$:

$$\begin{aligned} \mathbb{E}[\nabla_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t})|\mathcal{F}_t] &= \mathbb{E}[diag(\delta(\hat{\mathbf{Y}}_{t+h,t}-\mathbf{Y}_{t+h}))|\mathcal{F}_t] \\ &= diag(f_{t+h|t}(\hat{\mathbf{Y}}_{t+h,t})) \end{aligned}$$

where we have used the property of the Dirac Function that $\mathbb{E}_X[\delta(\gamma-X)] = \int_{-\infty}^{\infty} f_X(x)\delta(\gamma-x)dx = f_X(\gamma)$. As $diag(f_t(\hat{\mathbf{Y}}_{t+h,t})$ is a diagonal matrix we know that the eigenvalues are simply the diagonal entries. Therefore, by Assumption 3.B, $f_{t+h|t}(\hat{Y}_{t+h,t}) > 0\,\forall h$, we have shown that $\mathbb{E}[\nabla_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t})|\mathcal{F}_t]$ is positive definite.

For $\lambda > 1$,

$$\begin{aligned} \mathbb{E}[\nabla_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t})|\mathcal{F}_t] &= \lambda(\lambda-1)\mathbb{E}[diag(\kappa)|\mathcal{F}_t] + \lambda\mathbb{E}[diag(\nu)|\mathcal{F}_t] \\ &= \lambda(\lambda-1)\mathbb{E}[diag(\kappa)|\mathcal{F}_t], \end{aligned}$$

as $\mathbb{E}[\delta(\hat{Y}_{t+h,t} - Y_{t+h})|Y_{t+h} - \hat{Y}_{t+h,t}|^{\lambda-1}|\mathcal{F}_t] = 0\ \forall h$. Again, as $f_h(\boldsymbol{\alpha}_0) + (1 - 2f_h(\boldsymbol{\alpha}_0))1\{Y_{t+1} - \hat{Y}_{t+1,t} < 0\} > 0\,\forall\alpha \in \Theta$, $\forall h$ and $\lambda(\lambda-1) > 0\, for\, \lambda > 1$ we have that all the eigenvalues of $\mathbb{E}[\nabla_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t})|\mathcal{F}_t]$ are strictly positive. Hence, $\mathbb{E}[\nabla_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathcal{L}(\hat{\mathbf{Y}}_{t+H,t})|\mathcal{F}_t]$ is also positive definite for $\lambda > 1$. □

Proof of Proposition 2: As $\mathbf{W}$ is positive definite we know that $Q(\boldsymbol{\alpha}) = \mathbb{E}[h_{t+H}(\boldsymbol{\alpha})]^T\mathbf{W}\mathbb{E}[h_{t+H}(\boldsymbol{\alpha})]$, $h_{t+H}(\boldsymbol{\alpha}) = \mathbf{Z}_t \otimes diag(|\mathbf{Y}_{t+H} - \hat{\mathbf{Y}}^*_{t+H,t}|^{\lambda-1})(1\{\mathbf{Y}_{t+H} - \hat{\mathbf{Y}}^*_{t+H,t} < 0\} - f(\boldsymbol{\alpha}))$, is a strictly convex function and therefore admits a unique minimum in Θ. Hence, $\boldsymbol{\alpha}_0$ is uniquely defined as $\nabla_{\boldsymbol{\alpha}}Q(\boldsymbol{\alpha}_0) = 0$. □

Proof of Proposition 4: Proposition 2 show that $\boldsymbol{\alpha}_0$ is identified as the parameter value that minimises the quadratic form $Q(\boldsymbol{\alpha}) = \mathbb{E}[h_{t+H}(\boldsymbol{\alpha})]^T\mathbf{W}\mathbb{E}[h_{t+H}(\boldsymbol{\alpha})]$ where $h_{t+H}(\boldsymbol{\alpha}) = \mathbf{Z}_t \otimes diag[|\mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t}|^{\lambda-1}](1\{\mathbf{Y}_{t+H} - \mathbf{Y}^*_{t+H,t} < 0\} - f(\boldsymbol{\alpha}))$. The estimator, $\hat{\boldsymbol{\alpha}}_T$, is defined as the argument that minimizes the sample equivalent, $\hat{Q}_T(\boldsymbol{\alpha}) = \left(\frac{1}{T}\sum \hat{h}_{t+H}(\boldsymbol{\alpha})\right)^T \hat{\mathbf{W}}_T \left(\frac{1}{T}\sum \hat{h}_{t+H}(\boldsymbol{\alpha})\right)^T$, where we recommend setting $\hat{\mathbf{W}}_T = \mathbf{S}^{-1}$ as defined in equation 1.3.3. Initially, we will show that $\hat{Q}_T(\boldsymbol{\alpha})$ converges in probability to $Q(\boldsymbol{\alpha})$. Specifically, we will show that $\hat{\mathbf{h}}_T = \frac{1}{T}\sum \hat{h}_{t+H}(\boldsymbol{\alpha}) \xrightarrow{p} \mathbf{h}^* = \mathbb{E}[h_{t+H}(\boldsymbol{\alpha})]$ as $R, T \to \infty$. By assumption 4.B, we have that $\hat{\mathbf{W}}_T \xrightarrow{p} \mathbf{W}$. By the triangle inequality we get that

$$\left|\hat{\mathbf{h}}_T - \mathbf{h}^*\right| \leq \left|\hat{\mathbf{h}}_T - \hat{\mathbf{h}}\right| + \left|\hat{\mathbf{h}} - \mathbf{h}^*\right|$$

We first show that $\left|\hat{\mathbf{h}}_T - \hat{\mathbf{h}}\right| \xrightarrow{p} 0$ as $T \to \infty$ by using a law of large numbers for $\alpha - mixing$ processes. From Theorem 3.35 and 3.49 in White (2001), we know that measurable functions of strictly stationary and mixing processes are strictly stationary and mixing of the same size. Hence, by 4.A we have that $\{\mathbf{Z}_t \otimes diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)(1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t} < 0\} - f(\boldsymbol{\alpha}))\}$ is strictly stationary and $\alpha - mixing$ of size $-r/(r-1)$ with $r > 1$. In order to apply Corollary 3.48 in

White (2001), a law of large numbers for mixing processes, we need to ensure that

$$E\left[\left|\mathbf{Z}_t \otimes diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)(1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t}<0\}-f(\boldsymbol{\alpha},h))\right|^{r+\delta}\right]<\infty$$

for some $\delta>0$ and $r>1$. Initially, we can use that the modulus operator is multiplicative, $|a\times b|=|a|\times|b|$:

$$\begin{aligned} &E\left[\left|\mathbf{Z}_t \otimes diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)(1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t}<0\}-f(\boldsymbol{\alpha},h))\right|^{r+\delta}\right] = \\ &E\left[\left(|\mathbf{Z}_t|\times\left|diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)\right|\times\left|(1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t}<0\}-f(\boldsymbol{\alpha},h))\right|\right)^{r+\delta}\right]\end{aligned}$$

Furthermore, as $\left|(1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t}<0\}-f(\boldsymbol{\alpha},h))\right|\leq\boldsymbol{\iota}$ we get that

$$\begin{aligned} E\left[\left(|\mathbf{Z}_t|\otimes diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)\left|(1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t}<0\}-f(\boldsymbol{\alpha},h))\right|\right)^{r+\delta}\right] &\leq E\left[\left(|\mathbf{Z}_t|\otimes\left||\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right|\right)^{r+\delta}\right] \\ &= E\left[|\mathbf{Z}_t|^{r+\delta}\otimes|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{(\lambda-1)(r+\delta)}\right]\end{aligned}$$

By the Cauchy-Schwartz inequality, $E[xy]\leq E[x^2]^{1/2}E[y^2]^{1/2}$, we get that

$$E\left[|\mathbf{Z}_t|^{r+\delta}\otimes|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{(\lambda-1)(r+\delta)}\right]\leq\left\{E\left[|\mathbf{Z}_t|^{2(r+\delta)}\right]\otimes E\left[|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{2(\lambda-1)(r+\delta)}\right]\right\}^{1/2}$$

Now let $\delta=\frac{\varepsilon}{2}$

$$\left\{E\left[|\mathbf{Z}_t|^{2(r+\delta)}\right]\otimes E\left[|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{2(\lambda-1)(r+\delta)}\right]\right\}^{1/2} = \left\{E\left[|\mathbf{Z}_t|^{2r+\varepsilon)}\right]\otimes E\left[|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{(\lambda-1)(2r+\varepsilon)}\right]\right\}^{1/2}$$

Finally, by assumption 4.A, we get the necessary result:

$$E\left[\left|\mathbf{Z}_t \otimes diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)(1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t}<0\}-f(\boldsymbol{\alpha},h))\right|^{r+\delta}\right] \leq \{\Delta_1\otimes\Delta_2\}^{1/2}<\infty$$

Therefore, we can apply Corollary 3.48 in White (2001):

$$\begin{aligned}\frac{1}{T}\sum_{t=R+1}^{R+T}\mathbf{Z}_t\otimes diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)(1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t}<0\}-f(\alpha,h)) &\xrightarrow{p} \\ E[\mathbf{Z}_t\otimes diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)(1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t}<0\}-f(\boldsymbol{\alpha},h))] \quad as \quad T\to\infty\end{aligned}$$

Next we show that $\left|\hat{\mathbf{h}}-\mathbf{h}^*\right| \xrightarrow{p} 0$ as $R \to \infty$. Remember that

$$\begin{aligned}\left|\hat{\mathbf{h}}-\mathbf{h}^*\right| &= \Big|E[\mathbf{Z}_t \otimes diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)(1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t}<0\}-f(\boldsymbol{\alpha},h)) \\ &\quad -E[\mathbf{Z}_t \otimes diag\left(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1}\right)(1\{\boldsymbol{\epsilon}^*_{t+H,t}<0\}-f(\boldsymbol{\alpha},h))]\Big|\end{aligned}$$

Since$|\cdot|$ is a convex function we have by Jensen's inequality, $f(E[X]) \leq E[f(X)]$, that

$$\begin{aligned}\left|\hat{\mathbf{h}}-\mathbf{h}^*\right| &\leq E\Big[|\mathbf{Z}_t \otimes diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)(1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t}<0\}-f(\boldsymbol{\alpha},h)) \\ &\quad -\mathbf{Z}_t \otimes diag\left(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1}\right)(1\{\boldsymbol{\epsilon}^*_{t+H,t}<0\}-f(\boldsymbol{\alpha},h))|\Big] \\ &= E\Big[|\mathbf{Z}_t \otimes \Big(diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t}<0\} - diag\left(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1}\right)1\{\boldsymbol{\epsilon}^*_{t+H,t}<0\}\Big) \\ &\quad +\mathbf{Z}_t \otimes diag\left(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1}-|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)f(\boldsymbol{\alpha},h)|\Big]\end{aligned}$$

By the Triangular inequality, $|x+y| \leq |x|+|y|$ we get that

$$\begin{aligned}\left|\hat{\mathbf{h}}-\mathbf{h}^*\right| &\leq E\Big[|\mathbf{Z}_t \otimes \Big(diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t}<0\} - diag\left(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1}\right)1\{\boldsymbol{\epsilon}^*_{t+H,t}<0\}\Big)| \\ &\quad +|\mathbf{Z}_t \otimes diag\left(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1}-|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)f(\boldsymbol{\alpha},h)|\Big] \\ &= E\Big[|\mathbf{Z}_t| \otimes \Big|\Big(diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t}<0\} - diag\left(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1}\right)1\{\boldsymbol{\epsilon}^*_{t+H,t}<0\}\Big)\Big| \\ &\quad +|\mathbf{Z}_t| \otimes \Big|diag\left(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1}-|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)f(\boldsymbol{\alpha},h)\Big|\Big]\end{aligned}$$

where we used that the modulus operator is multiplicative in the second equality. Again, by the Cauchy-Schwartz inequality, $E[xy] \leq E[x^2]^{1/2}E[y^2]^{1/2}$, we get that

$$\begin{aligned}\left|\hat{\mathbf{h}}-\mathbf{h}^*\right| &\leq \left(E\left[|\mathbf{Z}_t|^2\right] \otimes E\left[\Big|\Big(diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t}<0\} - diag\left(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1}\right)1\{\boldsymbol{\epsilon}^*_{t+H,t}<0\}\Big)\Big|^2\right]\right)^{1/2} \\ &\quad +\left(E\left[|\mathbf{Z}_t|^2\right] \otimes E\left[\Big|diag\left(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1}-|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)\Big|^2 f(\boldsymbol{\alpha},h)^2\right]\right)^{1/2}\end{aligned}$$

Finally, by Assumption 4.D, and the continuous mapping theorem, we get that:

$$\begin{aligned}\left|\hat{\mathbf{h}}-\mathbf{h}^*\right| &\leq \left(E\left[|\mathbf{Z}_t|^2\right] \otimes E\left[\Big|\Big(diag\left(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1}\right)1\{\boldsymbol{\epsilon}^*_{t+H,t}<0\} - diag\left(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1}\right)1\{\boldsymbol{\epsilon}^*_{t+H,t}<0\}\Big)\Big|^2\right]\right)^{1/2} \\ &\quad +\left(E\left[|\mathbf{Z}_t|^2\right] \otimes E\left[\Big|diag\left(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1}-|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1}\right)\Big|^2 f(\boldsymbol{\alpha},h)^2\right]\right)^{1/2} \to 0 \ as \ R \to \infty.\end{aligned}$$

Therefore, $\frac{1}{T}\sum \hat{h}_{t+H}(\boldsymbol{\alpha}) \xrightarrow{p} \mathbb{E}[h_{t+H}(\boldsymbol{\alpha})]$ and, by the continuous mapping theorem, we have that $\hat{Q}_T(\boldsymbol{\alpha}) \xrightarrow{p} Q_0(\boldsymbol{\alpha})$. As the maximum of the limit is not necessarily equal to the limit of the maximum, we also need to ensure that $\hat{Q}_T(\boldsymbol{\alpha})$ converges uniformly to $Q(\boldsymbol{\alpha})$. This is achieved by employing Andrews' (2002) Theorem 1 which is valid under assumption 3.A (compact) and 4.C (Stochastic Equicontinuity). Therefore, we have shown that $\hat{\boldsymbol{\alpha}}_T \xrightarrow{p} \boldsymbol{\alpha}_0$ *as* $T, R \to \infty$. □

Proof of Proposition 5: To develop the asymptotic distribution of the estimator in proposition 5, we require an asymptotically valid closed form solution for $T^{1/2}(\hat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}_0)$. If $f(\boldsymbol{\alpha}) = \Pi\boldsymbol{\alpha}$, where $rank[\Pi] = k$, this is straightforward as lemma 3 show that there exist a closed for expression for $\hat{\boldsymbol{\alpha}}_T$ and therefore for $T^{1/2}(\hat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}_0)$. In the more general case, we need to rely on the mean value theorem. This theorem involves the gradient of $h_{t+H}(\boldsymbol{\alpha})$ and so it is necessary to impose assumption 5. The mean value theorem implies that

$$\hat{\mathbf{h}}_T(\hat{\boldsymbol{\alpha}}_T) = \hat{\mathbf{h}}_T(\boldsymbol{\alpha}_0) + \mathbf{m}_T(\tilde{\boldsymbol{\alpha}})(\hat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}_0)$$

where $\hat{\mathbf{h}}_T(\boldsymbol{\alpha}) = \frac{1}{T}\sum_{t=R+1}^{R+T} \mathbf{Z}_t \otimes diag(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1})(1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t} < 0\} - f(\boldsymbol{\alpha}))$,
$\hat{\mathbf{m}}_T(\boldsymbol{\alpha}) = \frac{1}{T}\sum_{t=1+R}^{R+T} \mathbf{Z}_t \otimes diag(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1})\frac{\partial f(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\alpha}} = diag(\boldsymbol{\gamma})\boldsymbol{\alpha}_0 + diag(\boldsymbol{\iota} - \boldsymbol{\gamma})\hat{\boldsymbol{\alpha}}_T$.
Premultiplication by $\hat{\mathbf{m}}_T(\hat{\alpha}_T)'\mathbf{W}_T$ yields:

$$\hat{\mathbf{m}}_T(\hat{\boldsymbol{\alpha}}_T)'\mathbf{W}_T\hat{\mathbf{h}}_T(\hat{\boldsymbol{\alpha}}_T) = \hat{\mathbf{m}}_T(\hat{\boldsymbol{\alpha}}_T)'\mathbf{W}_T\hat{\mathbf{h}}_T(\boldsymbol{\alpha}_0) + \hat{\mathbf{m}}_T(\hat{\boldsymbol{\alpha}}_T)'\mathbf{W}_T\hat{\mathbf{m}}_T(\tilde{\boldsymbol{\alpha}})(\hat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}_0).$$

Equation 1.3.3 imply that the left hand side is zero, and rearranging yieldd

$$\hat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}_0 = -\left(\hat{\mathbf{m}}_T(\hat{\boldsymbol{\alpha}}_T)'\mathbf{W}_T\hat{\mathbf{m}}_T(\tilde{\boldsymbol{\alpha}})\right)^{-1}\hat{\mathbf{m}}_T(\hat{\boldsymbol{\alpha}}_T)'\mathbf{W}_T\hat{\mathbf{h}}_T(\boldsymbol{\alpha}_0).$$

We get the desired expression by multiplying by $\sqrt{T}$

$$\sqrt{T}\,(\hat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}_0) = -\left(\hat{\mathbf{m}}_T(\hat{\boldsymbol{\alpha}}_T)'\mathbf{W}_T\hat{\mathbf{m}}_T(\tilde{\boldsymbol{\alpha}})\right)^{-1}\hat{\mathbf{m}}_T(\hat{\alpha}_T)'\mathbf{W}_T\left\{\sqrt{T}\hat{\mathbf{h}}_T(\boldsymbol{\alpha}_0)\right\}$$

By proposition 4 we know that $\hat{\boldsymbol{\alpha}}_T \to \boldsymbol{\alpha}_0$ as $T \to \infty$. Importantly, this implies that $\tilde{\boldsymbol{\alpha}} \to \boldsymbol{\alpha}_0$ as $T \to \infty$. By 5.A, we also know that

$$\begin{aligned} \hat{\mathbf{m}}_T(\boldsymbol{\alpha}) &= \frac{1}{T}\sum_{t=1+R}^{R+T} \mathbf{Z}_t \otimes diag(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1})\frac{\partial f(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \\ &\to \mathbb{E}\left[\mathbf{Z}_t \otimes diag(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda-1})\frac{\partial f(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right] \ as\ T, R \to \infty \\ &= \mathbf{m}(\boldsymbol{\alpha}_0) \end{aligned}$$

Finally, we use a CLT for $\alpha - mixing$ processes, Theorem 5.20 in White (2001), to derive the asymptotic distribution of $\sqrt{T}\hat{\mathbf{h}}_T(\boldsymbol{\alpha}_0)$. By Assumption 5.C, we know that $\{\mathbf{Z}_t \otimes diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)(1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t} < 0\} - f(\boldsymbol{\alpha}, h))\}$ is a strictly stationary and $\alpha - mixing$ of size $-r/(r-2)$ with $r > 2$, as Theorem 3.35 and 3.49 in White (2001) tells us that measurable functions of strictly stationary and mixing processes are strictly stationary and mixing of the same size. Furthermore, by 5.C we have that

$$\mathbb{E}\left[\left|\mathbf{Z}_t \otimes diag\left(|\hat{\boldsymbol{\epsilon}}^*_{t+H,t}|^{\lambda-1}\right)(1\{\hat{\boldsymbol{\epsilon}}^*_{t+H,t} < 0\} - f(\boldsymbol{\alpha}, h))\right|^r\right] <\leq \infty$$

for some $r > 2$ such that we can apply Theorem 5.20 in White (2001). Hence,

$$\begin{aligned} \sqrt{T}\hat{\mathbf{h}}_T(\boldsymbol{\alpha}_0) &= T^{1/2}\frac{1}{T}\sum_{t=1}^{T}[\mathbf{Z}_t \otimes diag(|\boldsymbol{\epsilon}^*_{t+H,t}|^{p_0-1})(1\{\boldsymbol{\epsilon}^*_{t+H,t} < 0\} - \boldsymbol{\alpha}_0)] \\ &\overset{\mathcal{L}}{\to} \mathcal{N}(0, \mathbf{S}(\boldsymbol{\alpha}_0)) \end{aligned}$$

where

$$\begin{aligned} \mathbf{S}(\boldsymbol{\alpha}_0) &= \lim_{T\to\infty} Var[T^{-1/2}\hat{\mathbf{h}}_T(\boldsymbol{\alpha}_0)] \\ &= \boldsymbol{\Gamma}_0(\boldsymbol{\alpha}_0) + \sum_{j=1}^{H-1}(\boldsymbol{\Gamma}_j(\boldsymbol{\alpha}_0)^* + \boldsymbol{\Gamma}_j(\boldsymbol{\alpha}_0)^{*T}) \end{aligned}$$

$$\boldsymbol{\Gamma}_0(\boldsymbol{\alpha}_0) = \mathbb{E}\left[\mathbf{Z}_t\mathbf{Z}_t^T \otimes \left(diag(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda_0-1})(1\{\boldsymbol{\epsilon}^*_{t+H,t} < 0\} - \mathbf{f}(\boldsymbol{\alpha}_0))\right)\left(diag(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda_0-1})(1\{\boldsymbol{\epsilon}^*_{t+H,t} < 0\} - \mathbf{f}(\boldsymbol{\alpha}_0))\right)^T\right]$$

$$\begin{aligned} \boldsymbol{\Gamma}_j(\boldsymbol{\alpha}_0)^* = \mathbb{E}\Big[&\mathbf{Z}_t\mathbf{Z}_{t+j}^T \otimes \left\{\mathbf{1}^j_{H\times H}\left(diag(|\boldsymbol{\epsilon}^*_{t+H,t}|^{\lambda_0-1})(1\{\boldsymbol{\epsilon}^*_{t+H,t} < 0\} - \mathbf{f}(\boldsymbol{\alpha}_0))\right)\right\} \times \\ &\left(diag(|\boldsymbol{\epsilon}^*_{t+H+j,t+j}|^{\lambda_0-1})(1\{\boldsymbol{\epsilon}^*_{t+H+j,t+j} < 0\} - \mathbf{f}(\boldsymbol{\alpha}_0))\right)^T\Big] \end{aligned}$$

where $\mathbf{1}^{j}_{H\times H} = diag(\boldsymbol{\iota}^{j}_{H}),\ \boldsymbol{\iota}^{j}_{H} = (0, 0, ..., 1, 1, ..., 1)^{T}$.The reason why the asymptotic variance has this particular form is that first n rationality conditions are martingale difference sequences and are therefore serially uncorrelated through time. However, the other $(H-1)n$ rationality conditions are not and this creates the particular autocorrelation terms, $\boldsymbol{\Gamma}_{j}(\boldsymbol{\alpha}_{o})^{*}$. Combining the two results we get that

$$\begin{aligned}\sqrt{T}\,(\hat{\boldsymbol{\alpha}}_{T} - \boldsymbol{\alpha}_{0}) &\overset{\mathcal{L}}{\to} \left(\mathbf{m}(\boldsymbol{\alpha}_{0})^{T}\mathbf{S}^{-1}\mathbf{m}(\boldsymbol{\alpha}_{0})\right)^{-1}\mathbf{m}(\boldsymbol{\alpha}_{0})\mathbf{S}^{-1}\mathcal{N}(0, \mathbf{S}) \\ &\overset{\mathcal{L}}{=} \mathcal{N}\left(0, \left(\mathbf{m}(\boldsymbol{\alpha}_{0})^{T}\mathbf{S}^{-1}\mathbf{m}(\boldsymbol{\alpha}_{0})\right)^{-1}\right)\end{aligned}$$

and proposition 5 follows. □

Appendix C

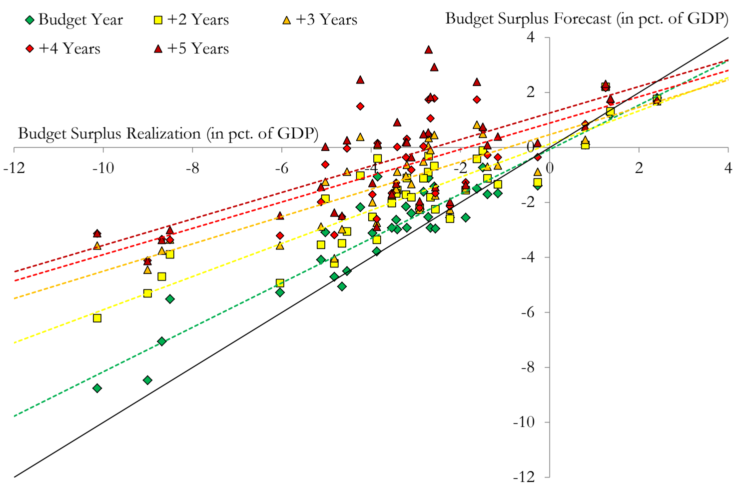


Figure 1.7: *Mincer-Zarnowitz regressions based on U.S. Administration budget surplus forecast, 1975-2012. The dotted lines depict the OLS fitted line and the black solid line depict the 45 degree line.*

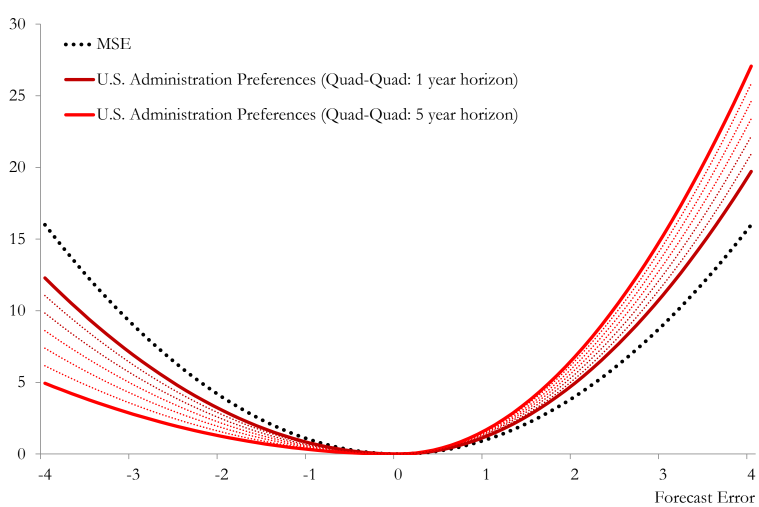


Figure 1.8: *Estimated loss function for the U.S. Administration based on single period Quad-Quad loss function specification, h=1,...,5. The estimation is based on the following instruments* $\mathbf{V}_t = [1, Y_t]^T$ *, 1975-2012.*

Appendix C (continued)

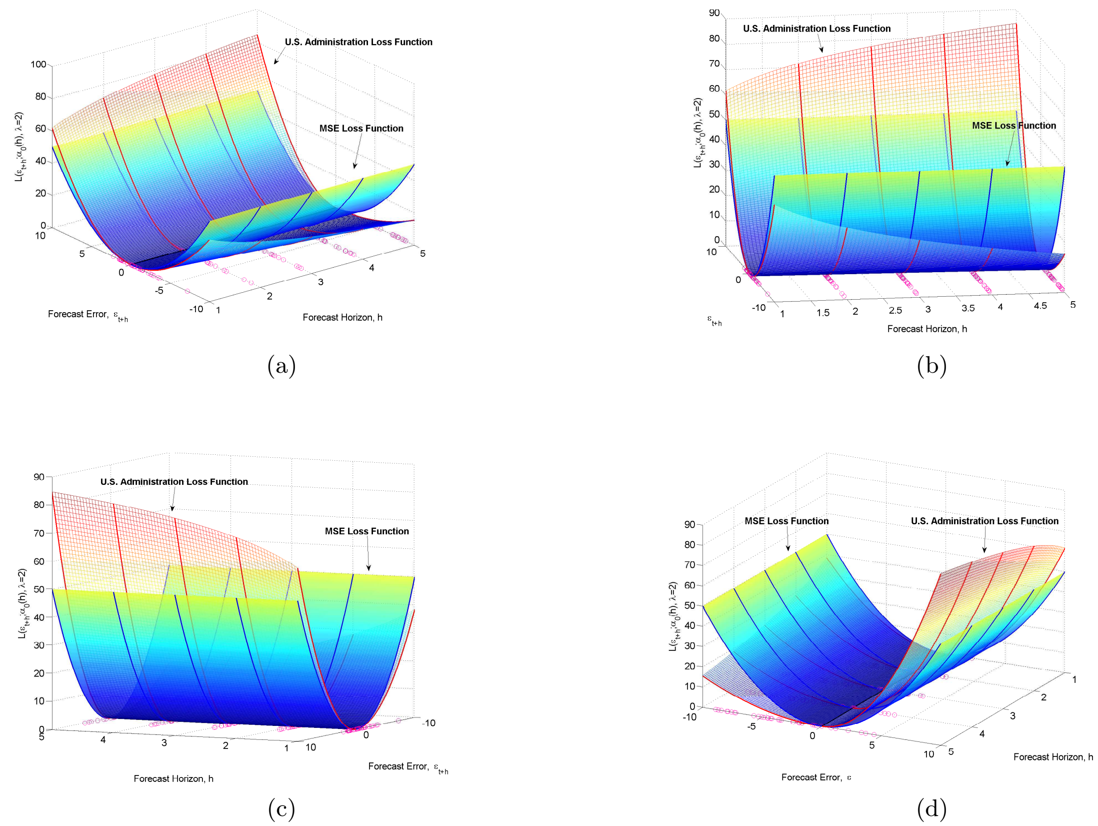


Figure 1.9: *Estimated U.S. Administration Loss Function based on 5-year term structure of forecasts for 1975-2012.*

1.8 Bibliography

Alesina, A. (1987), 'Macroeconomic policy in a two-party system as a repeated game', *The Quarterly Journal of Economics* **102**(3), 651–678.

Alesina, A., Roubini, N. & Cohen, G. (1997), *Political cycles and the macroeconomy*, the MIT Press.

Besley, T. & Case, A. (1995), 'Does electoral accountability affect economic policy choices? evidence from gubernatorial term limits', *The Quarterly Journal of Economics* **110**(3), 769–798.

Bohn, F. (2011), Getting around constitutional limits to fiscal policy: A model of political forecast and budget cycles. October 2011, Working Paper.

Boylan, R. (2008), 'Political distortions in state forecasts', *Public Choice* **136**(3), 411–427.

Brück, T. & Stephan, A. (2006), 'Do eurozone countries cheat with their budget deficit forecasts?', *Kyklos* **59**(1), 3–15.

Capistran, C. (2007), Optimality tests for multi-horizon forecasts. Working Paper 2007-14, Banco de Mexico.

Christoffersen, P. & Diebold, F. (1997), 'Optimal prediction under asymmetric loss', *Econometric Theory* **13**, 808–817.

Diebold, F. & Lopez, J. (1996), Forecast evaluation and combination, Technical report, National Bureau of Economic Research.

Elliott, G., Komunjer, I. & Timmermann, A. (2005), 'Estimation and testing of forecast rationality under flexible loss', *Review of Economic Studies* **72**(4), 1107–1125.

Elliott, G., Komunjer, I. & Timmermann, A. (2008), 'Biases in macroeconomic forecasts: irrationality or asymmetric loss?', *Journal of the European Economic Association* **6**(1), 122–157.

Frankel, J. (2011), 'Over-optimism in forecasts by official budget agencies and its implications', *Oxford Review of Economic Policy* **27**(4), 536–562.

Frankel, J. & Schreger, J. (2012), Over-optimistic official forecasts in the eurozone and fiscal rules, Technical report, National Bureau of Economic Research.

Granger, C. (1969), 'Prediction with a generalized cost of error function', *Journal of the Operational Research Society* **20**(2), 199–207.

Granger, C. W. J. & Newbold, P. (1986), *Forecasting Economic Time Series*, New York: Academic Press.

Granger, C. W. & Machina, M. J. (2006), 'Forecasting and decision theory', *Handbook of economic forecasting* **1**, 81–98.

Granger, C. W. & Pesaran, M. H. (2000), 'Economic and statistical measures of forecast accuracy', *Journal of Forecasting* **19**(7), 537–560.

Hibbs Jr, D. A. (1977), 'Political parties and macroeconomic policy', *American Political Science Review* **71**(4), 1467–1487.

Kliesen, K. & Thornton, D. (2012), 'How good are the governments deficit and debt projections and should we care?', *Federal Reserve Bank of St. Louis Review* **94**(1), 21–39.

Komunjer, I. & Owyang, M. T. (2012), 'Multivariate forecast evaluation and rationality testing', *Review of Economics and Statistics* **94**(4), 1066–1080.

MacRae, C. D. (1977), 'A political model of the business cycle', *Journal of Political Economy* **85**(2), pp. 239–263.

Mincer, J. A. & Zarnowitz, V. (1969), The evaluation of economic forecasts, *in* 'Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance', NBER, pp. 1–46.

Nordhaus, W. (1975), 'The political business cycle', *The Review of Economic Studies* **42**(2), 169–190.

Patton, A. J. & Timmermann, A. (2007*a*), 'Properties of optimal forecasts under asymmetric loss and nonlinearity', *Journal of Econometrics* **140**(2), 884–918.

Patton, A. & Timmermann, A. (2007*b*), 'Testing forecast optimality under unknown loss', *Journal of the American Statistical Association* **102**(480), 1172–1184.

Patton, A. & Timmermann, A. (2012), 'Forecast rationality tests based on multi-horizon bounds', *Journal of Business & Economic Statistics* **30**(1), 1–17.

Rogoff, K. (1990), 'Equilibrium political budget cycles', *The American Economic Review* **80**(1), pp. 21–36.

Rogoff, K. & Sibert, A. (1988), 'Elections and macroeconomic policy cycles', *The Review of Economic Studies* **55**(1), pp. 1–16.

Shi, M. & Svensson, J. (2006), 'Political budget cycles: Do they differ across countries and why?', *Journal of Public Economics* **90**(8), 1367–1389.

Varian, H. R. (1975), 'A bayesian approach to real estate assessment', *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage* **1**, 195–208.

West, K. D., Edison, H. J. & Cho, D. (1993), 'A utility-based comparison of some models of exchange rate volatility', *Journal of International Economics* **35**(1), 23–45.

White, H. (2001), *Asymptotic theory for econometricians*, Academic press New York.

Zellner, A. (1986), 'Bayesian estimation and prediction using asymmetric loss functions', *Journal of the American Statistical Association* **81**(394), 446–451.

Chapter 2

Monitoring Systemic Risk: Contagion, Risk Factors and Dynamic Policy Thresholds

Successful implementation of macroprudential policy is contingent on the ability to identify and estimate systemic risk in real time. In this paper systemic risk *is defined as the conditional probability of a systemic banking crisis (as defined by Reinhart & Rogoff 2011). This conditional probability is modelled in a fixed effect binary response panel-model framework that allows for cross-sectional dependence (e.g. due to contagion effects). In the empirical application we identify several risk factors and it is shown that the level of systemic risk contains a predictable component which varies through time. Furthermore, we illustrate how the forecasts of systemic risk map into dynamic policy thresholds in this framework. Finally, by conducting a pseudo out-of-sample exercise we find that the systemic risk estimates provided reliable early-warning signals ahead of the recent financial crisis for several economies.*

KEYWORDS: Systemic Risk, Macroprudential Policy, Binary Response Panel Model
JOURNAL OF ECONOMIC LITERATURE CODES: C23; C54; G18.

2.1 Introduction

The financial crisis in 2007-09 and the following global economic recession have highlighted the importance of a macroprudential policy framework which seeks to limit systemic financial risk. While there is still no consensus on how to implement macroprudential policy it is clear that successful implementation is contingent on establishing robust methods for monitoring systemic risk.[1] This current paper takes a step towards achieving this goal. Systemic risk assessment in real time is a challenging task due to the intrinsically unpredictable nature of systemic financial risk. However, as this study shows, in a fixed effect binary response model framework that allows for contagion effects that systemic risk does contain a component which varies in a predictable way through time and that modelling this component can potentially improve policy decisions.

In this paper, *systemic risk* is defined as the conditional probability of a systemic banking crisis; and we are interested in modelling and forecasting this (potentially) time-varying probability. If different systemic banking crises differ completely in terms of their underlying causes, triggers and economic impact, the conditional crisis probability will be unpredictable. However, as illustrated in section 2.4, systemic banking crises appear to share many commonalities. For example, banking crises are often preceded by prolonged periods of high credit growth and tend to occur when the banking sector is highly leveraged.

Systemic risk can be characterized by both cross-sectional and time-related dimensions (e.g. Bandt, Hartmann & Peydró 2009). The cross-sectional dimension concerns how risks are correlated across financial institutions at a given point in time as a result of direct and indirect linkages across institutions and prevailing default conditions. The time-series dimension concerns the evolution of systemic risk over time due to changes in the macroeconomic environment. This includes changes in the default cycle, changes in financial market conditions and the potential build-up of financial imbalances such as asset and credit

[1] In November 2010, G20 leaders called on the International Monetary Fund, the Bank for International Settlements and the Financial Stability Board to do further work on monitoring and regulating systemic risk. Their report supported the view that monitoring systemic risk is an important element of successful macroprudential policy.

market bubbles. The focus in this paper is on the time dimension of systemic risk, although the empirical analysis includes a variable that proxies for the strength of interconnectedness between financial institutions.

This paper makes the following contributions to the literature on systemic risk assessment: firstly, it provides conditions for consistency and asymptotic normality of the model parameters in a fixed effect binary response model framework that allows for cross-sectional dependence. Secondly, it applies this approach to a large panel of 68 advanced and emerging economies in order to identify leading indicators of systemic risk. While Demirgüç-Kunt & Detragiache (1998*a*) study the determinants of banking crises, the purpose of this paper is to evaluate whether systemic risk can be monitored in real time. Consequently, it employs a model structure that is purely dynamic such that the systemic risk forecasts are based solely on information available in real time. In addition, the parameters are estimated by a partial likelihood approach with fixed effects that provide consistent model parameter estimates under more general conditions than the random effect approach that has been used in other studies (e.g. Demirgüç-Kunt & Detragiache (1998*a*) and Wong, Wong & Leung (2010)). Thirdly, this chapter shows how to derive dynamic risk-factor thresholds in the binary response model framework. The threshold of a single risk factor is dynamic in the sense that it depends on the value of the other risk factors, and it is argued that this approach has some advantages relative to static thresholds based on the signal extraction approach.[2] Finally, we perform a pseudo out-of-sample analysis for the period 2001-2010 in order to assess whether the risk factors provided early warning signals ahead of the recent financial crisis.
Based on the empirical analysis, we reach the following main conclusions:

1. *Systemic risk*, as defined here, does appear to be predictable in real time. In particular, the following risk factors are identified: banking sector leverage, equity price growth, the credit-to-GDP gap, real effective exchange rate appreciation, changes in banks' lending premiums and the degree of banks' interconnectedness as measured by the

[2] The signal extraction approach was popularized in this literature by Kaminsky & Reinhart (1999). See section 2.5 for more details.

ratio of non-core to core bank liabilities. There is also some evidence which suggests that house-price growth increases systemic risk, but the effect is not statistically significant at conventional significance levels.

2. There exist significant contagion effects between economies. When an economy with a large financial sector is experiencing a systemic banking crisis, the systemic risk forecasts in other economies increase significantly.

3. Rapid credit growth in a country is often associated with a higher level of systemic risk. However, as highlighted in a recent IMF report (2011), a boom in credit can also reflect a healthy market response to expected future productivity gains as a result of new technology, new resources or institutional improvements. Indeed, many episodes of credit booms were not followed by a systemic banking crisis nor any other material instability. Thus, it is critical that a policy maker is able to distinguish between these two scenarios when implementing economic policy. We find empirical evidence which suggests that credit growth increases systemic risk considerably more when accompanied by high-equity price growth. Therefore, we argue that the evolution in equity prices can be useful for identifying a healthy credit expansion.

4. In a crisis signalling exercise, we find that the binary response model approach outperforms the popular signal extraction approach in terms of type I and type II errors.

5. Based on a model specification with credit-to-GDP growth, banking sector leverage and equity price growth, we carefully evaluated the optimal credit-to-GDP growth threshold. Contrary to the signal extraction approach the optimal threshold is not static but depends on the value of the other risk factors. For example, the threshold is around 10 percent if equity prices have decreased by 10 percent and banking sector leverage is around 130 percent but only around 0 percent if equity prices have grown by 20 percent and banking sector leverage is 160 percent. In comparison, the signal extraction method leads to a (static) credit-to-GDP growth threshold of 4.9 percent based on the same data sample.

6. In the out-of-sample analysis we find that the systemic risk factors generally provided informative signals in many countries. Based on an in-sample calibration, around 50-80 percent of the crises were correctly identified in real time without constructing too many false signals. In particular, a monitoring model based on credit-to-GDP growth and banking sector leverage signalled early warning signals ahead of the U.S. sub-prime crisis in 2007.

The rest of the paper is organized as follows. Section 2 contains a brief literature overview. Section 3 presents the modelling approach and provide conditions for consistency and asymptotic normality of the model parameters in the presence of cross-sectional dependence. Section 4 presents the empirical results. Section 5 illustrates how the estimated models can be used for monitoring purposes and how to derive optimal risk-factor thresholds. Finally, section 6 concludes.

2.2 Related Literature

The purpose of this section is to provide a brief overview of the literature. A more comprehensive review can be found in Bell & Pain (2000), Gaytán & Johnson (2002) and Demirgüç-Kunt & Detragiache (2005). Understanding the causes and triggers of systemic banking crises has long been a core interest of regulators, central bankers and academics, and there is a vast literature on the subject.[3] There are generally two approaches to this question that can be found in the empirical literature: one, the signal extraction approach and, two, the econometric approach.

The signal extraction approach was popularized in this field by Kaminsky & Reinhart (1999) who focused on the phenomenon of the 'twin crises', namely the simultaneous occurrence of currency and banking crises. They document the incidence of currency, banking and twin crises in a sample of twenty industrial and emerging countries from 1970 to 1995. The paper

[3] See, for example, Davis & Karim (2008), Goldstein, Kaminsky & Reinhart (2000), González-Hermosillo (1996, 1999) and IMF (1998, 2011).

looked at 16 potential indicators and found that the real exchange rate, stock prices and the ratio of public sector deficits to GDP were the most useful indicators. Later, Borio & Lowe (2002), Borio & Drehmann (2009), among others, employed the same approach to identify leading indicators of systemic banking crises. They found that the credit-to-GDP gap was the most useful indicator of systemic risk.

Demirgüç-Kunt & Detragiache (1998*a*) used an econometric approach to study the determinants of systemic banking crisis during the period of 1980-1994. Their empirical results indicated that systemic banking distress was associated with a weak macroeconomic environment in which there was low economic growth, high inflation and high real interest rates. In addition, they found that balance of payments crises were also associated with banking crises. Other studies such as Demirgüç-Kunt & Detragiache (1998*a*,*b*, 2000, 2002, 2005), Hutchison & McDill (1999), Domac & Martinez Peria (2003) and Wong, Wong & Leung (2010) followed a similar econometric approach.

This paper employs an econometric approach, but it differentiates itself from previous studies in the following dimensions. First, the purpose of this study is to evaluate whether systemic risk can be monitored in real time rather than identifying the determinants of systemic banking crisis. In some sense this is a less ambitious goal because we do not necessarily need to worry about interpreting coefficients as representing causal effects, but it forces us to employ a purely dynamic model-specification such that the systemic risk forecasts are solely based on information available *before* the event occurs. The model specification in Demirgüç-Kunt & Detragiache (1998*a*) is not appropriate for systemic risk monitoring since it utilizes contemporaneous economic variables which are not available before a crisis occurs. Second, contrary to the studies mentioned, this study employs a fixed effect estimation approach to estimate the model parameters. This estimator is consistent under weaker conditions than a random effects estimator since it does not require the risk factors to be independent of the unobserved country fixed effects. If the independence assumption is true, the two estimators should converge towards the same (true) parameter value as the number of observations goes to infinity. Interestingly, we find that the fixed effect estimator and the random effects estimator generate quite different parameter estimates. Since both estimators are consistent under the independence assumption, this suggests that the risk factors and the unobserved country fixed effects are not independent and therefore that the random effects estimator is inconsistent in the model specification.[4] This study also provides conditions for consistency and asymptotic normality of the model parameters in the presence of cross-sectional dependence. This is an important result because systemic banking crises are potentially dependent in the cross section due to contemporaneous contagion effects.

The popularity of the signal extraction approach is partly due to its easily interpretable thresholds. However, one potential drawback of this approach is that the thresholds are static. This implies that if the threshold of credit growth is 5 percent then a crisis signal would be issued if credit growth exceeds this level regardless of what was happening with the other factors that might impact the level of systemic risk. This is problematic as one could argue that a boom in credit can also reflect a healthy market response to expected future productivity gains that are the result of new technology, new resources or institutional improvements (IMF 2011). This paper derives dynamic risk-factor thresholds that are based on the binary response model framework. This approach allows for a more realistic environment where appropriate policy thresholds depend on the state of the economy via several other risk factors.

2.3 Modelling Approach

In this paper *systemic risk* is defined as the conditional probability of a systemic banking crisis,

$$Systemic\,Risk_{i,t-1} \equiv Pr(y_{i,t} = 1\,|\mathcal{F}_{t-1}),$$

[4] A potential drawback of the fixed effects estimator, in this binary response panel data model, is that the beta estimator is unreliable if the time-series dimension T_i is small such that the unobserved country fixed effects, the $\alpha_i's$, are estimated imprecisely. This is known as the incidental parameter problem (Neyman & Scott (1948) and Lancaster (2000)). However, since the time dimension of the data is relatively large in this study, 10-40 years, this is a minor problem. For example, Heckman (1981) found that for N=100 and T=8 the bias appeared to be of order 10%.

where $y_{i,t}$ is a binary systemic banking crisis variable (as defined by Reinhart & Rogoff (2011)), i denotes the country, t denotes the time period and $\mathcal{F}_t$ denotes a $\sigma - algebra$ (to be defined below) that represents the information set at time t. The econometric analysis is based on the assumption that the conditional probability varies over time in a systematic way. More specifically, we model the binary systemic banking crisis variable in a fixed effect panel model framework,

$$
\begin{aligned}
y^*_{i,t} &= \alpha_i + \mathbf{x}^T_{i,t-h}\boldsymbol{\beta} + \epsilon_{i,t}\,,\; \epsilon_{i,t}|\mathcal{F}_{t-1} \sim i.i.d\, F(0,\sigma^2) \\
y_{i,t} &= \mathbf{1}\{\alpha_i + \mathbf{x}^T_{i,t-h}\boldsymbol{\beta} + \epsilon_{i,t} > 0\}\,,\; i = 1,...,N \text{ and } t = 1,...,T,
\end{aligned}
$$

where $\mathcal{F}_t \equiv \sigma\{\mathbf{Z}_s,\, s \le t\}$, $\mathbf{Z}_s = \left(y_{1,s},...,y_{N,s},\mathbf{x}^T_{1,s},...,\mathbf{x}^T_{N,s}\right)^T$, $y^*_{i,t}$ is an unobserved latent variable, $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_N)^T \in \mathcal{A} \subset \mathbb{R}^N$ is a $N \times 1$ vector of time invariant fixed effects, $\boldsymbol{\beta} \in \mathcal{B} \subset \mathbb{R}^k$ is a $k \times 1$ vector of unknown parameters, $\mathbf{1}\{\cdot\}$ is an indicator function that takes the value unity if the condition is true and zero otherwise, and $\mathbf{x}_{i,t-1}$ is a $k \times 1$ vector of random variables. Finally, $\epsilon_{i,t}$ is an unobserved idiosyncratic error term with zero conditional mean, known conditional variance σ^2 and a condition distribution given by $F(0,\sigma^2)$. At first glance, it appears to be a strong assumption to assume that the conditional mean and variance of the unobserved error term is known. However, both assumptions are innocent normalizations. First, suppose the variance of $\epsilon_{i,t}$ is scaled by an unrestricted parameter γ. The latent regression will be $y^*_{i,t} = \alpha_i + \mathbf{x}^T_{i,t-h}\boldsymbol{\beta} + \gamma^{0.5}\epsilon_{i,t}$, but $\gamma^{-0.5}y^*_{i,t} = \gamma^{-0.5}\alpha_i + \gamma^{-0.5}\mathbf{x}^T_{i,t-h}\boldsymbol{\beta} + \epsilon_{i,t}$ is the same model with the same data. The observed data will be unchanged, as $y_{i,t}$ is 1 or 0 depending only on the sign of $y^*_{i,t}$. This implies that there is no information about γ in the data, so it cannot be estimated. Second, the assumption of zero mean is also innocent as the model includes a time invariant fixed effect, α_i. In order to illustrate this let δ be a non-zero threshold so that the probability that $y_{i,t}$ equals one is $Pr(\alpha_i + \mathbf{x}^T_{i,t-h}\boldsymbol{\beta} + \epsilon_{i,t} > \delta\,|\mathcal{F}_{t-h}) = Pr(\tilde{\alpha}_i + \mathbf{x}^T_{i,t-h}\boldsymbol{\beta} + \epsilon_{i,t} > 0\,|\mathcal{F}_{t-h})$, where $\tilde{\alpha}_i = \alpha_i - \delta$. Therefore, as α_i is unknown, $\tilde{\alpha}_i$ remains an unknown parameter.

We will also assume that the distribution of $\epsilon_{i,t}$ is symmetric around zero, such that $Pr(\epsilon_{i,t} > -\alpha_i - \mathbf{x}^T_{i,t-h}\boldsymbol{\beta}\,|\mathcal{F}_{t-h}) = Pr(\epsilon_{i,t} \le \alpha_i + \mathbf{x}^T_{i,t-h}\boldsymbol{\beta}\,|\mathcal{F}_{t-1})$, although this assumption can be relaxed. The conditional probability, or equivalently the level of systemic risk, can then be written as,

$$\begin{aligned} Pr(y_{i,t} = 1\,|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i) &= Pr(\epsilon_{i,t} > -\alpha_i - \mathbf{x}_{i,t-h}^T\boldsymbol{\beta}\,|\mathcal{F}_{t-1}), \\ &= G(\alpha_i + \mathbf{x}_{i,t-h}^T\boldsymbol{\beta}) \end{aligned}$$

where $G(\alpha_i + \mathbf{x}_{i,t-h}^T\boldsymbol{\beta}) = \int_{-\infty}^{\alpha_i+\mathbf{x}_{i,t-1}^T\boldsymbol{\beta}} f_{\epsilon_{i,t}}(z|\mathcal{F}_{t-1})dz$, known as the link function, is the cumulative distribution function of the idiosyncratic error term. Therefore, in this latent regression model the form of the link function is determined by the conditional distribution function of the error term, $F(0, \sigma^2)$. This implies that the link function has the following nice properties, $\underset{x\to\infty}{lim}\, G(x) = 1$ and $\underset{x\to-\infty}{lim}\, G(x) = 0$, such that the conditional probability is bounded, $Pr(y_{i,t} = 1\,|\mathcal{F}_{t-1}) \in [0, 1]$.

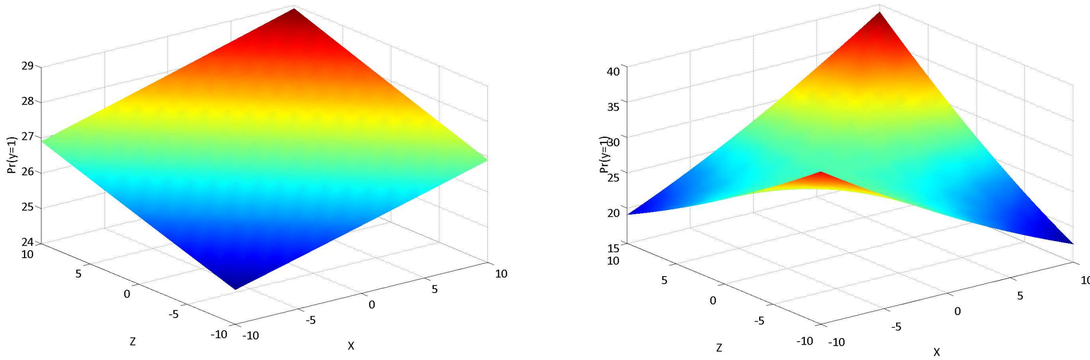
As the main purpose of this chapter is to monitor systemic risk, rather than to identify the determinants of systemic banking crises, the model structure is dynamic such that all the risk factors are known (measurable) h periods (years) in advance. This is crucial as it gives a policy maker time to react to a crisis signal.

2.3.1 Marginal Effects

Contrary to a standard linear regression model, the marginal effects are generally not constant in this binary, fixed effect model framework. The marginal effect of an incremental increase in $x_{ij,t-h}$, an element of $\mathbf{x}_{i,t-h}$, is given by:

$$\frac{\partial Pr(y_{i,t} = 1|\mathcal{F}_{t-1}; \alpha_i, \boldsymbol{\beta})}{\partial x_{ij,t-h}} = \left.\frac{\partial G'(z)}{\partial z}\right|_{z=\alpha_i+\mathbf{x}_{i,t-h}^T\boldsymbol{\beta}} \times \beta_j.$$

We define the standardized marginal effect as $(\partial Pr(y_{i,t} = 1|\mathbf{x}_{i,t-h}; \alpha_i, \boldsymbol{\beta})/\partial x_{ij,t-h}) \times Var[x_{j,t}]^{0.5}$. It measures the approximate increase in systemic risk due to a standard deviation increase in a risk factor. It is clear that the marginal effect depends on the value of all the risk factors and on the value of the country fixed effect, α_i. Therefore, the model structure implies that there is an implicit interaction effect between all the risk indicators. The marginal effect of an increase in a risk indicator is



(a) No explicit interaction term (b) Explicit interaction term

Figure 2.1: *Binary Response Model Structure. The conditional probabilities are computed as: (a)* $Pr(y = 1|x, z) = \Lambda(-1+0.5\%x+0.5\%z)$ *and (b)* $Pr(y = 1|x, z) = \Lambda(-1+0.5\%x+0.5\%xz)$.

higher, all else equal, if the other risk indicators are also high.[5] This is illustrated in figure 2.1a. According to the theoretical model predictions in Allen & Gale (2000*a*), the marginal increase in systemic risk following an asset price bubble is higher if it is accompanied by a credit expansion. The binary response model structure is consistent with this dynamic.

It is also worth pointing out that introducing an explicit interaction term in the model, $x_{i,t-h} \times z_{i,t-h}$, will generate some perverse model dynamics as illustrated in figure 2.1b.

For example, if x takes a negative value then the marginal effect of z can change sign. This is not usually an appealing property and we will avoid this model structure in the paper. Finally, note also that the country fixed effects are treated as parameters to be estimated. This approach is more robust than a random effects estimator where consistency requires that the risk indicators and the country fixed effects are stochastically independent.

[5] To be precise, the marginal effect of a risk indicator is higher if the other risk indicators are high, provided that the other risk factors have positive beta coefficients and that the conditional probability is smaller than 50% for the logit or probit specification. In the application in this chapter, this will generally be true.

2.3.2 Parameter Estimation

The model structure implies that $y_{i,t}|\mathcal{F}_{t-1} \sim i.i.d\, Bernoulli\left(G(\alpha_i + \mathbf{x}_{i,t-h}^T \boldsymbol{\beta})\right)$, such that the conditional density function of $y_{i,t}$ is given by

$$f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i) = G(\alpha_i + \mathbf{x}_{i,t-h}^T \boldsymbol{\beta})^{y_{i,t}}[1 - G(\alpha_i + \mathbf{x}_{i,t-h}^T \boldsymbol{\beta})]^{1-y_{i,t}}\,.$$

The parameters, $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T \in \Theta \subset \mathbb{R}^{kN}$ will be estimated by a partial likelihood approach. What distinguishes this approach from a maximum likelihood approach is that we do not model the joint density function for all countries at time t, given by $f(y_{1,t}, ..., y_{N,t}|\mathcal{F}_{t-1}; \boldsymbol{\theta})$. Generally, by Sklar's theorem, we have that

$$f(\mathbf{y}_t|\mathcal{F}_{t-1}; \boldsymbol{\theta}) = c(F_1(y_{1,t}), ..., F_N(y_{N,t})) \prod_{i=1}^{N} f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i),$$

where $c(\cdot)$ is the copula function that determines the dependence structure.[6] Therefore, knowledge of all the sub-models does not deliver knowledge of the joint density unless the individual components are conditionally independent. In many microeconometric applications it is reasonable to assume that the observations are conditionally independent in the cross section (e.g. individual $i's$ decision to enter the labour force is independent of individual $j's$ decision). However, the independence assumption is less reasonable in the application in this paper. For example, if there are contagion effects across countries the probability of a systemic banking crisis occuring in one economy might increase if another, closely related, economy is experiencing a systemic banking crisis. One way to mitigate this potential problem would be to include a time varying, country specific, contemporaneous contagion-effect variable as an explanatory variable. However, such a variable would not be measurable before a crisis occurs, which means it is therefore not feasible to use for monitoring purposes. In order to address this potential issue, we rely on a partial likelihood approach to estimate the parameters. We define the partial log likelihood for each time

[6] Note that while Sklar's theorem does not necessarily require that the marginal CDF's are continuous (e.g. Rüschendorf 2009), the copula function is not unique when the distribution is discrete (or has discrete parts).

period t as

$$
\begin{aligned}
l_t(\boldsymbol{\theta}; \mathbf{y}_t, \mathbf{X}_{t-h}) &= \sum_{i=1}^{N} log[f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i)] \\
&= \sum_{i=1}^{N} y_{i,t} log[G(\alpha_i + \mathbf{x}_{i,t-h}^T \boldsymbol{\beta})] + (1 - y_{i,t}) log[1 - G(\alpha_i - \mathbf{x}_{i,t-h}^T \boldsymbol{\beta})],
\end{aligned}
$$

which is the sum of likelihoods across countries. While we do not impose that the observations are conditionally independent in the cross section, the model structure implies that $y_{i,t}|\mathcal{F}_{t-1}$ are conditionally independent through time for all i. The partial likelihood estimator is defined as:

$$
\hat{\boldsymbol{\theta}}_{PL} = \underset{\boldsymbol{\theta} \in \Theta}{arg\,max} \left\{ \frac{1}{T} \sum_{t=1}^{T} l_t(\boldsymbol{\theta}; \mathbf{y}_t, \mathbf{X}_{t-h}) \right\}. \tag{2.3.1}
$$

This estimator is potentially inefficient as we are ignoring the potential dependence in the data across countries. What makes this partial likelihood approach work is that, by the Kullback-Leibler information inequality, $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)'$ maximizes $\mathbb{E}\left[log[f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i)]\right]$ over Θ for each i, so $\boldsymbol{\theta}_0$ also maximizes the sum of these over i. Note that in order to guarantee identification we need to verify that $\boldsymbol{\theta}_0$ uniquely maximizes $\mathbb{E}\left[log[f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i)]\right]$ for each i. We will now analyze the asymptotic properties of the estimator.

2.3.3 Asymptotic Properties of the Partial Likelihood Estimator

Since the model setup implies that $y_{it}|\mathcal{F}_{t-1} \sim i.i.d$ through time we have that $\frac{1}{T}\sum_{t=1}^{T} log[f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i) \xrightarrow{p} \mathbb{E}\left[log[f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i)]\right]$ as $T \to \infty$, for all i, such that $\frac{1}{T}\sum_{t=1}^{T} l_t(\boldsymbol{\theta}; \mathbf{y}_t, \mathbf{X}_{t-h}) \xrightarrow{p} \mathbb{E}\left[\sum_{i=1}^{N} log[f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i)]\right]$. However, as the partial likelihood estimator, as defined in 2.3.1, is an extremum estimator, it is not sufficient to ensure that the estimated objective function converges pointwise to the true objective function in order to obtain consistency as the maximum of a limit is not necessarily equal to the limit of the maximum. A sufficient condition is uniform convergence (see e.g. Newey & McFadden

(1994)):

$$\max_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{T} l_t(\boldsymbol{\theta};\mathbf{y}_t,\mathbf{X}_{t-1}) - \mathbb{E}[l_t(\boldsymbol{\theta}_0;\mathbf{y}_t,\mathbf{X}_{t-1})]\right| \xrightarrow{p} 0\,.$$

Following theorem 1 in Andrews (1992), sufficient conditions for uniform convergence is that Θ is compact and that $\frac{1}{T}\sum_{t=1}^{T} l_t(\boldsymbol{\theta};\mathbf{y}_t,\mathbf{X}_{t-1})$ is stochastically equicontinuous on Θ.[7] Therefore, under assumptions A.1-A.5, stated in appendix 2.7.4, we can prove the following consistency proposition.

Proposition 1. *Let* $\hat{\boldsymbol{\theta}}_{PL}$ *be the partial likelihood estimator defined in* (2.3.1). *If assumptions (A1)-(A5) hold then:* $\hat{\boldsymbol{\theta}}_{PL} \xrightarrow{p} \boldsymbol{\theta}_0$ *as* $T\to\infty$.

Proof. See appendix 2.7.4. □

It is worth pointing out that nonlinear fixed effects models, in a standard microeconometric setting, generally suffer from the incidental parameters problem analysed by Neyman & Scott (1948) and Lancaster (2000). The problem arises because the asymptotics in a standard microeconometric framework are derived for $N\to\infty$ and fixed T.[8] As such, since α_i is only estimated by T observations, the fixed effects are not consistently estimated as $N\to\infty$. Furthermore, as the estimator of $\boldsymbol{\beta}$ is a function of the estimators of the fixed effects, $\boldsymbol{\beta}$ is also not consistently estimated. These two points are clearly illustrated by looking at the gradient of the partial log likelihood function:

$$\begin{aligned}
\frac{\partial T^{-1}\sum_{t=1}^{T} l_t(\boldsymbol{\beta},\boldsymbol{\alpha};\mathbf{y}_t,\mathbf{X}_t)}{\boldsymbol{\beta}} &= \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{G'(\alpha_i+\mathbf{x}_{i,t-h}^T\boldsymbol{\beta})(y_{i,t}-G(\alpha_i+\mathbf{x}_{i,t-h}^T\boldsymbol{\beta})}{G(\alpha_i+\mathbf{x}_{i,t-h}^T\boldsymbol{\beta})(1-G(\alpha_i+\mathbf{x}_{i,t-h}^T\boldsymbol{\beta}))}\mathbf{x}_{i,t-h}\\
\frac{\partial T^{-1}\sum_{t=1}^{T} l_t(\boldsymbol{\beta},\boldsymbol{\alpha};\mathbf{y}_t,\mathbf{X}_t)}{\alpha_1} &= \frac{1}{T}\sum_{t=1}^{T}\frac{G'(\alpha_1+\mathbf{x}_{1,t-h}^T\boldsymbol{\beta})(y_{1,t}-G(\alpha_1+\mathbf{x}_{1,t-h}^T\boldsymbol{\beta})}{G(\alpha_1+\mathbf{x}_{1,t-h}^T\boldsymbol{\beta})(1-G(\alpha_1+\mathbf{x}_{1,t-h}^T\boldsymbol{\beta}))}\\
&\vdots\\
\frac{\partial T^{-1}\sum_{t=1}^{T} l_t(\boldsymbol{\beta},\boldsymbol{\alpha};\mathbf{y}_t,\mathbf{X}_t)}{\alpha_N} &= \frac{1}{T}\sum_{t=1}^{T}\frac{G'(\alpha_N+\mathbf{x}_{N,t-h}^T\boldsymbol{\beta})(y_{N,t}-G(\alpha_N+\mathbf{x}_{N,t-h}^T\boldsymbol{\beta})}{G(\alpha_N+\mathbf{x}_{N,t-h}^T\boldsymbol{\beta})(1-G(\alpha_N+\mathbf{x}_{N,t-h}^T\boldsymbol{\beta}))}
\end{aligned}$$

where the first k equations depends on $\boldsymbol{\alpha}$ and the last N equations clearly do not converge as $N\to\infty$. However, the incidental parameter problem disappears in this framework as the

[7]**Definition:** $G_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} l_t(\boldsymbol{\theta};\mathbf{y}_t,\mathbf{X}_{t-h})$ is stochastically equicontinuous on Θ if $\forall\epsilon>0\,\exists\delta>0$ such that $\lim_{T\to\infty} Pr(\sup_{\theta\in\Theta}\sup_{\theta'\in B(\theta,\delta)}|G_T(\theta')-G_T(\theta)|>\epsilon)<\delta$.

[8]Moreover, not only is T fixed, it is quite small in many microeconometric applications.

asymptotics are derived for $T \to \infty$. Importantly, this is a fairly reasonable approximation in this application as the data covers the period 1970-2010. The incidental parameter problem is therefore likely to be of minor importance. For example, Heckman (1981) found that for N=100 and T=8, in a probit fixed effect model the bias appeared to be of order 10 percent.

In order to derive the asymptotic distribution of the partial likelihood estimator we need to make some additional technical assumptions; specifically, A.6-A.9 as stated in 2.7.4. For example, we need to assume that $\boldsymbol{\theta}_0 \in \mathring{\Theta}$ and that $l_t(\boldsymbol{\theta}; \mathbf{y}_t, \mathbf{X}_{t-h})$ is twice differentiable on $\mathring{\Theta}$. Then, one can prove the following proposition:

Proposition 2. *Let $\hat{\boldsymbol{\theta}}_{PL}$ be the partial likelihood estimator defined in* (2.3.1)*. If the assumptions of proposition 1 and (A6-A9) hold then*

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{H}_0^{-1}\mathbf{J}_0\mathbf{H}_0^{-1}),$$

where $\mathbf{J}_0$ and $\mathbf{H}_0$ are defined as

$$\mathbf{J}_0 = \mathbb{E}\left[\begin{pmatrix} \sum_{i=1}^{N} \frac{[y_{i,t}-G(a_{i,t}(\boldsymbol{\theta}_i))]G'(a_{i,t}(\boldsymbol{\theta}_i))}{G(a_{i,t}(\boldsymbol{\theta}_i))(1-G(a_{i,t}(\boldsymbol{\theta}_i)))}\mathbf{x}_{i,t-h} \\ \frac{(y_{1,t}-G(a_{1,t}(\boldsymbol{\theta}_1)))G'(a_{1,t}(\boldsymbol{\theta}_1))}{G(a_{1,t}(\boldsymbol{\theta}_1))(1-G(a_{1,t}(\boldsymbol{\theta}_1)))} \\ \vdots \\ \frac{(y_{N,t}-G(a_{N,t}(\boldsymbol{\theta}_N)))G'(a_{N,t}(\boldsymbol{\theta}_N))}{G(a_{N,t}(\boldsymbol{\theta}_N))(1-G(a_{N,t}(\boldsymbol{\theta}_N)))} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{N} \frac{[y_{i,t}-G(a_{i,t}(\boldsymbol{\theta}_i))]G'(a_{i,t}(\boldsymbol{\theta}_i))}{G(a_{i,t}(\boldsymbol{\theta}_i))(1-G(a_{i,t}(\boldsymbol{\theta}_i)))}\mathbf{x}_{i,t-h} \\ \frac{(y_{1,t}-G(a_{1,t}(\boldsymbol{\theta}_1)))G'(a_{1,t}(\boldsymbol{\theta}_1))}{G(a_{1,t}(\boldsymbol{\theta}_1))(1-G(a_{1,t}(\boldsymbol{\theta}_1)))} \\ \vdots \\ \frac{(y_{N,t}-G(a_{N,t}(\boldsymbol{\theta}_N)))G'(a_{N,t}(\boldsymbol{\theta}_N))}{G(a_{N,t}(\boldsymbol{\theta}_N))(1-G(a_{N,t}(\boldsymbol{\theta}_N)))} \end{pmatrix}^T\right]$$

$$\mathbf{H}_0 = -\mathbb{E}\begin{bmatrix} \sum_{i=1}^{N} q_{i,t}(\boldsymbol{\theta}_i)\mathbf{x}_{i,t-h}\mathbf{x}_{i,t-h}^T & q_{1,t}(\boldsymbol{\theta}_i)\mathbf{x}_{1,t-h} & \cdots & q_{N,t}^0\mathbf{x}_{N,t-h} \\ q_{1,t}(\boldsymbol{\theta}_1)\mathbf{x}_{1,t-h}^T & q_{1,t}(\boldsymbol{\theta}_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ q_{N,t}(\boldsymbol{\theta}_N)\mathbf{x}_{N,t-h}^T & 0 & \cdots & q_{N,t}(\boldsymbol{\theta}_i) \end{bmatrix}$$

$$\begin{aligned} a_{i,t}(\boldsymbol{\theta}_i) &= \alpha_i + \mathbf{x}_{i,t-h}^T\boldsymbol{\beta} \\ q_{i,t}(\boldsymbol{\theta}_i) &= \left(\frac{[y - G(a_{i,t}(\boldsymbol{\theta}_i))]G'(a_{i,t}(\boldsymbol{\theta}_i)}{G(a_{i,t}(\boldsymbol{\theta}_i))(1 - G(a_{i,t}(\boldsymbol{\theta}_i)))}\right)^2. \end{aligned}$$

Proof. See appendix 2.7.4. □

Remark: If the binary systemic banking crises are conditionally uncorrelated in the cross section, $Cov[y_{i,t}, y_{j,t}|\mathcal{F}_{t-1}] = 0 \; \forall i \neq j$, the asymptotic variance simplifies to $avar(\hat{\boldsymbol{\theta}}_{PL}) = \mathbf{H}_0^{-1}$ as the cross products in $\mathbf{J}_0$ cancels out. This result is shown in appendix 2.7.4.

The consistency and asymptotic normality result of the partial likelihood estimator holds for a large class of link functions. In the application in this paper, we will follow a conventional approach and assume that $\epsilon_{i,t}|\mathcal{F}_{t-1} \sim i.i.d.\, \mathcal{N}(0,1)$ or $\epsilon_{i,t}|\mathcal{F}_{t-1} \sim i.i.d.\, Logistic\left(0, \frac{\pi^2}{3}\right)$. This choice leads to the probit or logit model respectively.

Example 1: Fixed Effect Logit Model

Suppose that $\epsilon_{i,t}|\mathcal{F}_{t-1} \sim i.i.d.\, Logistic\left(0, \frac{\pi^2}{3}\right)$ such that $G(\alpha_i + \mathbf{x}_{i,t-h}^T\boldsymbol{\beta}) = \Lambda(\alpha_i + \mathbf{x}_{i,t-h}^T\boldsymbol{\beta})$ where $\Lambda(x) = e^x/(1+e^x)$ is the cumulative distribution function for a standard logistic distribution. The identification condition (A.2) is satisfied if $\mathbb{E}[(1,\, \mathbf{x}_{i,t-h}^T)^T(1,\, \mathbf{x}_{i,t-h}^T)]$ is non-singular. Furthermore, the partial log-likelihood function, $\sum_{i=1}^N log[f(y_{i,t}|\mathcal{F}_{t-h};\boldsymbol{\beta},\alpha_i)] = y_{i,t}log[\Lambda(\alpha_i + \mathbf{x}_{i,t-h}^T\boldsymbol{\beta})] + (1-y_{i,t})log[1-\Lambda(\alpha_i - \mathbf{x}_{i,t-h}^T\boldsymbol{\beta}),$ satisfies (A.7) as it is a twice differentiable function in Θ. The asymptotic variance is given by

$$\mathbf{J}_0 = \mathbb{E}\left[\begin{pmatrix} \sum_{i=1}^N [y_{i,t} - \Lambda(a_{i,t}(\boldsymbol{\theta}_i))]\mathbf{x}_{i,t-h} \\ [y_{1,t} - \Lambda(a_{1,t}(\boldsymbol{\theta}_1))]\mathbf{x}_{1,t-h} \\ \vdots \\ [y_{N,t} - \Lambda(a_{N,t}(\boldsymbol{\theta}_N))]\mathbf{x}_{N,t-h} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^N [y_{i,t} - \Lambda(a_{i,t}(\boldsymbol{\theta}_i))]\mathbf{x}_{i,t-h} \\ [y_{1,t} - \Lambda(a_{1,t}(\boldsymbol{\theta}_1))]\mathbf{x}_{1,t-h} \\ \vdots \\ [y_{N,t} - \Lambda(a_{N,t}(\boldsymbol{\theta}_N))]\mathbf{x}_{N,t-h} \end{pmatrix}^T\right]$$

$$\mathbf{H}_0 = -\mathbb{E}\begin{bmatrix} \sum_{i=1}^N m_{i,t}(\boldsymbol{\theta}_i)\mathbf{x}_{i,t-h}\mathbf{x}_{i,t-h}^T & m_{1,t}(\boldsymbol{\theta}_1)\mathbf{x}_{1,t-h} & \cdots & m_{N,t}(\boldsymbol{\theta}_N)\mathbf{x}_{N,t-h} \\ m_{1,t}(\boldsymbol{\theta}_1)\mathbf{x}_{1,t-h}^T & m_{1,t}(\boldsymbol{\theta}_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ m_{N,t}(\boldsymbol{\theta}_N)\mathbf{x}_{N,t-h}^T & 0 & \cdots & m_{N,t}(\boldsymbol{\theta}_N) \end{bmatrix},$$

where $a_{i,t}(\boldsymbol{\theta}_i) = (\alpha_i + \mathbf{x}_{i,t-h}^T\boldsymbol{\beta}_0)$ and $m_{i,t}(\boldsymbol{\theta}_i) = \Lambda(a_{i,t}(\boldsymbol{\theta}_i))[1-\Lambda(a_{i,t}(\boldsymbol{\theta}_i))]$.

Example 2: Fixed Effect Probit Model

Alternatively, suppose that $\epsilon_{i,t}|\mathcal{F}_{t-1} \sim i.i.d.\mathcal{N}(0,1)$ such that $G(\alpha_i + \mathbf{x}_{i,t-h}^T\boldsymbol{\beta}) = \int_{-\infty}^{\alpha_i+\mathbf{x}_{i,t-h}^T\boldsymbol{\beta}} \phi(z|\mathcal{F}_{t-h})dz$ where $\phi(\cdot)$ denotes the standard normal density function. The identification condition (A.2) is again satisfied if $\mathbb{E}[(1\,\mathbf{x}_{i,t-h}^T)^T(1\,\mathbf{x}_{i,t-h}^T)]$ is non-singular. Furthermore, the partial log-likelihood function, $\sum_{i=1}^N log[f(y_{i,t}|\mathcal{F}_{t-h};\boldsymbol{\beta},\alpha_i)] = y_{i,t}log[\Phi(\alpha_i + \mathbf{x}_{i,t-h}^T\boldsymbol{\beta})] + (1-y_{i,t})log[1-\Phi(\alpha_i - \mathbf{x}_{i,t-h}^T\boldsymbol{\beta}),$ satisfies (A.7) as it is a twice differentiable function in Θ. The asymptotic variance is given by

$$\mathbf{J}_0 = \mathbb{E}\left[\begin{pmatrix} \sum_{i=1}^N \frac{[y_{i,t}-\Phi(a_{i,t}(\boldsymbol{\theta}_i))]\phi(a_{i,t}(\boldsymbol{\theta}_i))\mathbf{x}_{i,t-h}}{\Phi(a_{i,t}(\boldsymbol{\theta}_i))(1-\Phi(a_{i,t}(\boldsymbol{\theta}_i)))} \\ \frac{(y_{1,t}-\Phi(a_{1,t}(\boldsymbol{\theta}_1)))\phi(a_{1,t}(\boldsymbol{\theta}_1))\mathbf{x}_{1,t-h}}{\Phi(a_{1,t}(\boldsymbol{\theta}_1))(1-\Phi(a_{1,t}(\boldsymbol{\theta}_1)))} \\ \vdots \\ \frac{(y_{N,t}-\Phi(a_{N,t}(\boldsymbol{\theta}_N)))\phi(a_{N,t}(\boldsymbol{\theta}_N))\mathbf{x}_{N,t-h}}{\Phi(a_{N,t}(\boldsymbol{\theta}_N))(1-\Phi(a_{N,t}(\boldsymbol{\theta}_N)))} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^N \frac{[y_{i,t}-\Phi(a_{i,t}(\boldsymbol{\theta}_i))]\phi(a_{i,t}(\boldsymbol{\theta}_i))\mathbf{x}_{i,t-h}}{\Phi(a_{i,t}(\boldsymbol{\theta}_i))(1-\Phi(a_{i,t}(\boldsymbol{\theta}_i)))} \\ \frac{(y_{1,t}-\Phi(a_{1,t}(\boldsymbol{\theta}_1)))\phi(a_{1,t}(\boldsymbol{\theta}_1))\mathbf{x}_{1,t-h}}{\Phi(a_{1,t}(\boldsymbol{\theta}_1))(1-\Phi(a_{1,t}(\boldsymbol{\theta}_1)))} \\ \vdots \\ \frac{(y_{N,t}-\Phi(a_{N,t}(\boldsymbol{\theta}_N)))\phi(a_{N,t}(\boldsymbol{\theta}_N))\mathbf{x}_{N,t-h}}{\Phi(a_{N,t}(\boldsymbol{\theta}_N))(1-\Phi(a_{N,t}(\boldsymbol{\theta}_N)))} \end{pmatrix}\right]$$

$$\mathbf{H}_0 = -\mathbb{E}\begin{bmatrix} \sum_{i=1}^N \frac{\{\phi(a_{i,t}(\theta_i)\}^2\mathbf{x}_{i,t-h}\mathbf{x}_{i,t-h}^T}{\Phi(a_{i,t}(\theta_i))[1-\Phi(a_{i,t}(\theta_i))]} & \frac{\{\phi(a_{i,t}(\theta_i)\}^2\mathbf{x}_{i,t-h}}{\Phi(a_{i,t}(\theta_i))[1-\Phi(a_{i,t}(\theta_i))]} & \cdots & \frac{\{\phi(a_{i,t}(\theta_i)\}^2\mathbf{x}_{i,t-h}}{\Phi(a_{i,t}(\theta_i))[1-\Phi(a_{i,t}(\theta_i))]} \\ \frac{\{\phi(a_{i,t}(\theta_i)\}^2\mathbf{x}_{i,t-h}^T}{\Phi(a_{i,t}(\theta_i))[1-\Phi(a_{i,t}(\theta_i))]} & \frac{\{\phi(a_{i,t}(\theta_i)\}^2}{\Phi(a_{i,t}(\theta_i))[1-\Phi(a_{i,t}(\theta_i))]} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\{\phi(a_{i,t}(\theta_i)\}^2\mathbf{x}_{i,t-h}^T}{\Phi(a_{i,t}(\theta_i))[1-\Phi(a_{i,t}(\theta_i))]} & 0 & \cdots & \frac{\{\phi(a_{i,t}(\theta_i)\}^2}{\Phi(a_{i,t}(\theta_i))[1-\Phi(a_{i,t}(\theta_i))]} \end{bmatrix},$$

where Φ denotes the cumulative distribution function of a standard normal random variable.

Finally, note that $\mathbf{J}_0$ and $\mathbf{H}_0$ are unobservable. In order to compute standard errors we simply use the sample equivalents of the population moments:

$$\hat{\mathbf{J}}_0 = T^{-1}\sum_{t=1}^T\left[\left(\sum_{i=1}^N \frac{\partial log[f(y_{i,t}|\mathcal{F}_{t-h};\hat{\beta}_{PL},\hat{\alpha}_{i,PL})]}{\partial\boldsymbol{\theta}}\right)\left(\sum_{i=1}^N \frac{\partial log[f(y_{i,t}|\mathcal{F}_{t-h};\hat{\beta}_{PL},\hat{\alpha}_{i,PL})]}{\partial\boldsymbol{\theta}}\right)^T\right]$$

$$\hat{\mathbf{H}}_0 = T^{-1}\sum_{t=1}^T\sum_{i=1}^N \mathbb{E}\left[\frac{\partial^2 log[f(y_{i,t}|\mathcal{F}_{t-h};\hat{\beta}_{PL},\hat{\alpha}_{i,PL})]}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right].$$

More generally, it is feasible to model the conditional probability using more general link functions. For example, by simply assuming that $G(x) = x$, one gets the linear probability model (LPM): $Pr(y_{i,t} = 1|\mathbf{x}_{i,t-h};\alpha_i,\boldsymbol{\beta}) = \alpha_i + \mathbf{x}_{i,t-h}^T\boldsymbol{\beta}$. A drawback of this model is that the probability is not bounded between 0 and 1. We will primarily focus on the logit model specification, but the results are not sensitive to this choice as is illustrated in section 2.4.

After having shown that the partial likelihood estimator is consistent and asymptotically normal, we now turn our attention to the model specification.

2.3.4 Model Specification

The empirical analysis is based on an unbalanced annual panel of 68 advanced and emerging economies over the time period 1970-2010. Table 2.1 contains an overview of the economies. Due to the rare nature of systemic banking crisis it is advantageous to use a panel data approach. In this way one is able to exploit information from several time-series and get more reliable and precise parameter estimates of $\boldsymbol{\beta}$.[9] On the other hand, the drawback is that one is imposing the potential false restriction that $\boldsymbol{\beta}$ is common for all the countries.

For the dependent variable, $y_{i,t}$, we adopt the definition of a systemic banking crisis from Reinhart & Rogoff (2011). They define a banking crisis to be systemic if one of the following conditions is satisfied:

- There are bank runs that lead to closure, merger or takeover by the public sector of one or more financial institutions;
- Or, if there are no runs, but there are closures, mergers, takeovers, or large-scale government assistance of an important financial institution that marks the start of a string of similar outcomes for other financial institutions.

Laeven & Valencia (2008, 2010) employ a similar definition as illustrated in figure 2.2a. Interestingly, a larger number of countries were experiencing a systemic banking crisis in the mid-90's than compared with the recent 2007-09 financial crises. This is a bit surprising at first glance. However, after correcting for the relative sizes of the economies, the recent financial crisis is more severe as indicated by figure 2.2b.

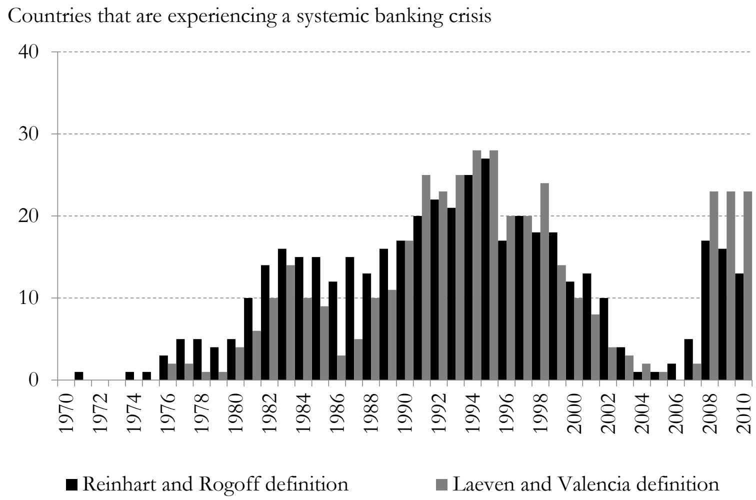
[9] Note that potential cross-sectional dependence in the systemic banking crises will reduce the efficiency gains from this panel data approach. However, even in the presence of cross-sectional dependence, this approach should lead to more precise parameter estimates.

	Advanced Economies:		
Austria	France	Japan	Singapore
Australia	Germany	South Korea	Spain
Belgium	Greece	Netherlands	Sweden
Canada	Iceland	New Zealand	Switzerland
Denmark	Ireland	Norway	United Kingdom
Finland	Italy	Portugal	United States
	Emerging Economies:		
Algeria	Dominican Republic	Malaysia	Romania
Angola	Ecuador	Mexico	Russia
Argentina	Egypt	Morocco	South Africa
Bolivia	El Salvador	Myanmar	Sri Lanka
Brazil	Ghana	Nicaragua	Thailand
Central African Republic	Guatemala	Nigeria	Tunisia
Chile	Honduras	Panama	Turkey
China	Hungary	Paraguay	Uruguay
Colombia	India	Peru	Venezuela
Costa Rica	Indonesia	Philippines	Zambia
Côte d'Ivoire	Kenya	Poland	Zimbabwe

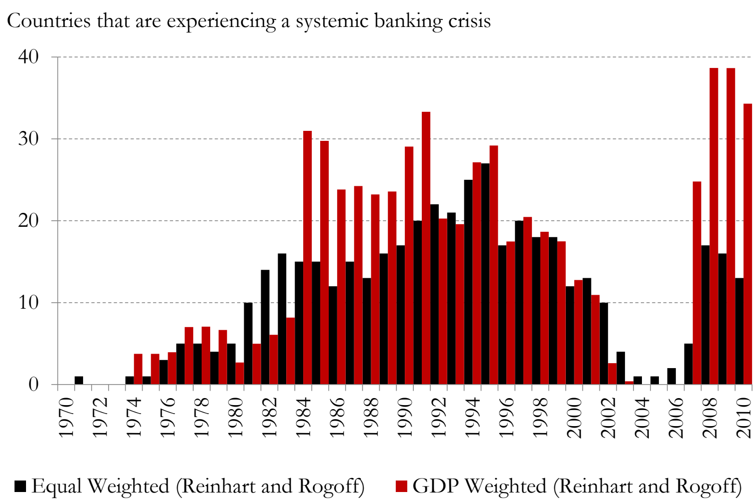
Table 2.1: Countries in Data Sample

As potential risk factors we focus on the following: private credit by deposit money banks as a percentage of GDP, asset prices, the banks' lending premium, the real effective exchange rate, banking sector leverage and bank interconnectedness as proxied by the ratio of non-core to core bank liabilities. We will motivate this choice in the following.

The real effective exchange rate is a macroeconomic variable which is related to a country's international trade competitiveness. The idea behind including this variable as a risk factor is that a deteriorating international trade competitiveness would be more likely to amplify an initial shock and create a systemic banking crisis. Therefore, we conjecture that real effective exchange rate appreciation will have a positive impact on a country's systemic risk.



(a) Equal Weighted



(b) GDP Weighted

Figure 2.2: *Systemic Banking Crises, 1970-2010. Details of the crises dates can be found in Table 2.10 in Appendix 2.7.3. For further details regarding the definitions of systemic banking crises see Reinhart and Rogoff (2010) and Laeven and Valencia (2008, 2010). The graph in figure (b) is computed as* $\sum_{i=1}^{N} \omega_i y_{i,t}$*, where the weights,* $\{\omega_i\}_{i=1}^{N}$*, are based on GDP in 2005, i.e.* $\omega_i = GDP_i^{2005} \cdot (N^{-1} \sum_{j=1}^{N} GDP_j^{2005})^{-1}$.

Asset prices are represented by equity and by house prices. These prices tend to move cyclically around a long-run trend over time. Therefore, high asset price inflation might indicate a higher possibility that prices may fall back to the trend level in the future, and price corrections like this are likely to happen when adverse economic shocks occur. Furthermore, banks are vulnerable to such asset price declines since sharp declines in collateral value and rising defaults on loans will deteriorate a bank's balance sheets.

According to the theoretical model presented in Allen & Gale (2000*a*), credit growth amplifies asset price inflation, thus generating a higher systemic risk. Empirically, Borio & Lowe (2002) and Borio & Drehmann (2009) also find that credit-to-GDP growth and the credit-to-GDP gap is associated with systemic banking crises. The credit-to-GDP gap is the difference between the current level and the long-term trend. The (equilibrium) trend level is estimated by a backward-looking Hodrick-Prescott filter (HP-filter) which is estimated recursively for each time period. When using the HP-filter one has to choose the smoothing parameter, λ. Following Drehmann, Borio & Tsatsaronis (2011), we set λ=1600 reflecting that financial cycles are approximately four times longer than standard business cycles.[10]

The banking sector is generally more vulnerable to adverse shocks when it is highly leveraged. Therefore, we also include a banking sector leverage variable as a potential risk indicator. We define *leverage* as private credit by deposit money banks as a percentage of demand, time and saving deposits in deposit money banks.[11]

Theoretically, even a single bank default can pose a threat to the financial system via its dependence to other financial institutions (Diamond & Dybvig (1983), and Allen & Gale (2000*b*)). In order to capture this effect we also include a measure of the degree of interconnectedness in the financial sector. All else equal, if banks are highly interconnected a single bank default is more likely to trigger a systemic banking crisis. Hahm, Shin & Shin (2011) argue that the ratio of non-core to core bank liabilities is related to the degree of financial interconnectedness and is signalling financial vulnerability. A bank's core liabilities consist of retail deposits from the household sector. This form of funding does not increase

[10]Hodrick & Prescott (1997) recommended setting the smoothing parameter λ to 100 for annual data. For robustness, we also tried this and other smoothing parameter values but it does not have any substantial impact on the results.

[11]Note that the definition of banking sector leverage differs slightly from the standard definition of leverage as total assets as a percentage of total equity.

systemic risk as it does not increase the dependence between banks. Non-core liabilities, on the other hand, consist of funding from other financial institutions. This type of funding is 'bad' in the sense that it increases the banks' interconnectedness and therefore increases the systemic risk in the economy. Therefore, we propose to use non-core liabilities, as a percentage of core liabilities, as a potential risk indicator. Following Hahm, Shin & Shin (2011), we adopt two alternative measures of non-core bank liabilities:

$$\begin{aligned} Non-Core\,1 &= Banks\,liabilities\,to\,the\,foreign\,sector \\ &\quad +Banks\,liabilities\,to\,the\,non-banking\,financial\,sector \end{aligned}$$

$$\begin{aligned} Non-Core\,2 &= Banks\,liabilities\,to\,the\,foreign\,sector \\ &\quad +(M3-M2) \end{aligned}$$

Both measures include bank liabilities to the foreign sector, which constitutes an important source of non-deposit wholesale funding for banks in emerging and developing economies. In addition, the Non-Core 1 definition adds bank liabilities to non-bank financial institutions such as insurance companies and pension funds, and the Non-Core 2 definition adds M3 - M2 as an additional component. We use broad money, M2, to measure core liabilities.

In order to evaluate whether there exists an intertemporal contagion effect between the economies we also try to include a contagion variable which is defined as $Contagion_{i,t} = \sum_{i \neq j} \omega_{j,t} y_{j,t}$, where $y_{j,t}$ is the binary systemic banking crisis variable for country j at time t, and $\omega_{j,t}$ is country $j's$ market capitalization, at time t, as a percentage of the world's market capitalization. If a systemic banking crisis in a large economy increases the level of systemic risk in other countries then one should observe a significant positive coefficient estimate for this variable.

Finally, we also include the change in the banks' lending premium. The *lending premium* is defined as the difference between the interest rate charged by banks on loans to prime private-sector customers minus the 'risk free' treasury bill interest rate at which short-term government securities are issued or traded. The purpose of including this variable is to capture time variation in conditional bank returns.

2.4 Estimation Results

This section presents the estimation results based on the binary response model framework discussed in section 2.3. Detailed definitions and sources of data are given in appendix 2.7.1.

To get an overview of the proposed risk indicators' abilities to detect systemic risk we initially estimated single variable model specifications. As a robustness check we considered three different models: logit, probit and the linear probability model (LPM). The estimation results are presented in tables 2.6 and 2.7 in appendix 2.7.2.

Based on these tables it is clear that the β coefficients vary between the three model specifications. This is not surprising as the link functions are different. What matters are the marginal effects and they are a function of both the model parameters and the link function. Figure 2.3 illustrates the standardized marginal effects for the three model specifications, i.e. the marginal effects multiplied by the standard deviation of the risk indicator, for the three model specifications. Based on this figure it is clear that the results are quite consistent across the three model specifications. This is comforting as it reveals that the results are not sensitive to the choice of link function.

In the following, we will focus on the logit model specification where the link function is given by $\Lambda(x) = \frac{e^x}{1+e^x}$. Figure 2.4 illustrates the β estimates and their corresponding confidence intervals for the logit specification. Based on figures 2.4a and 2.4b, it is clear that the credit-to-GDP growth and the credit-to-GDP gap have a positive significant effect on the level of systemic risk up to three years in advance. This is consistent with the findings in Borio & Lowe (2002) and Borio & Drehmann (2009) who also find that these variables are useful leading indicators of systemic risk. It is also interesting to note that the parameter estimates are similar for both advanced and emerging economies. One advantage of the binary response model approach, relative to the signalling extraction method, is that one is

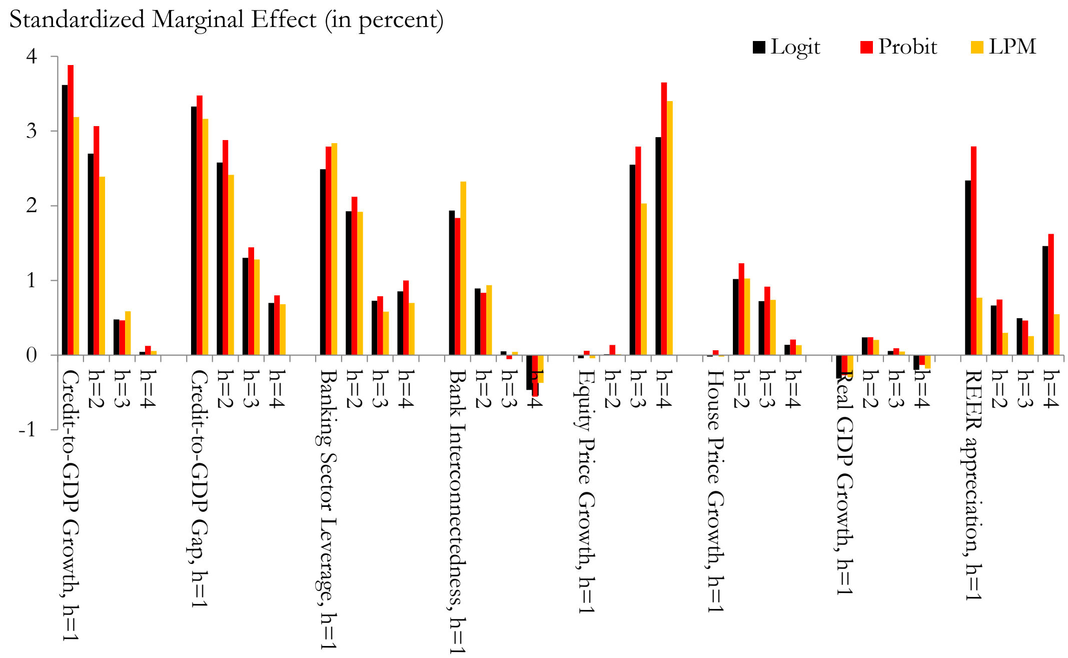


Figure 2.3: *Standardized Marginal Effects. The marginal effects for the logit and the probit model specifications are evaluated at the median country fixed effect and the median value of the risk indicator. The standardized marginal effect is defined as:* $(\partial Pr(y_{i,t} = 1|\mathbf{x}_{i,t-h}; \alpha_i, \boldsymbol{\beta})/\partial x_{ij,t-h}) \times Var[x_{j,t}]^{0.5}$. *It approximates the marginal increase in systemic risk due to one standard deviation increase in a risk indicator. Models with different lags (h) are estimated using the same data sample. See Table 2.6 and 2.7 for detailed estimation results.*

able to quantify the increase in systemic risk following an increase in a risk indicator. The results suggest that a one standard deviation increase in the credit-to-GDP gap, evaluated at the median country fixed effect and median risk-factor value, increases the systemic risk by around 3.3 percent, 2.6 percent and 1.3 percent the following three years for the logit model. This is illustrated in figure 2.3.

Based on figure 2.4c it is also clear that a leveraged banking sector is associated with higher systemic risk. This effect is significant up to two years in advance. For a median risk country, at a one year forecast horizon, a one standard deviation increase in banking leverage increases systemic risk by 2.5 percent based on the logit specification, cf. Figure 2.3.

Bank interconnectedness, as proxied by non-core to core bank liabilities, is also found to have a significant impact on the level of systemic risk at a one-year forecast horizon, figure 2.4d.

This is consistent with the theoretical model presented in Hahm, Shin & Shin (2011) where non-core funding increases the banks' interconnectedness and therefore increases the likelihood of a systemic banking crisis if a single bank defaults.

High asset-price inflation is also found to be systematically associated with systemic banking crises. For the single-predictor model specification, equity price inflation has a significant positive impact on systemic risk, three to four years ahead as seen in figure 2.4e. House price inflation also appears to increase systemic risk at a two-year forecast horizon, although the effect is not statistically significant at a five percent significance level. This is illustrated in figure 2.4f.

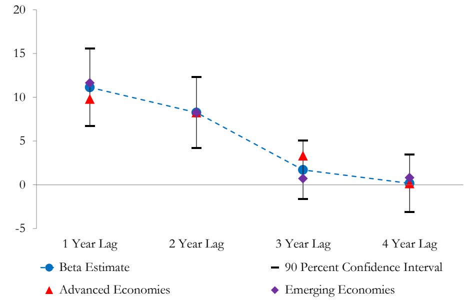
Deteriorating trade competitiveness, measured by a real effective exchange rate appreciation, is also found to increase systemic risk as seen in figure 2.4g. Not surprisingly, emerging economies appear to be more sensitive to this effect.

Real GDP growth does not appear to have any impact on systemic risk based on the single-factor logit model. However, it appears that the impact is asymmetric on emerging and advanced economies. Emerging economies tend to experience negative GDP growth leading up to a systemic banking crisis while advanced economies tend to experience positive GDP growth as seen in figure 2.4h.

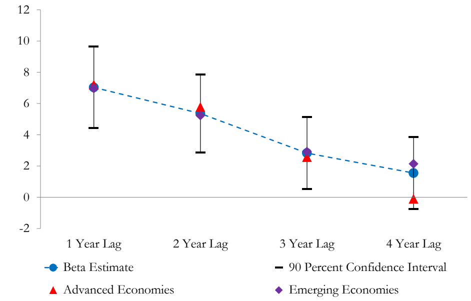
After the initial, single-factor analysis we now turn to a multivariate analysis where we allow the risk indicators to interact with each other implicitly. We focus again on the logit model specification, but the results are similar for the probit model specification, as indicated by figure 2.3. Table 2.2 presents the results for various model specifications.

The key findings are summarized as follows:

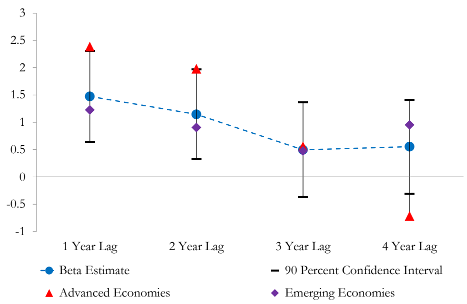
1. Combining banking-sector leverage, credit-to-GDP gap and equity-price inflation appears to provide a good representation of the data for both a one- and two-year forecast horizon. All the estimated coefficients are positive and significant at a 5



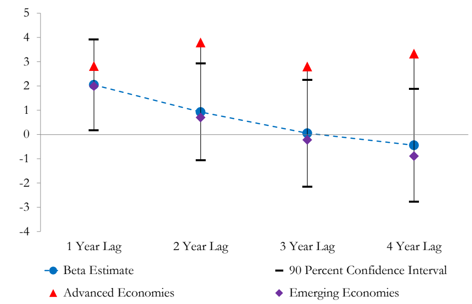
(a) Credit-to-GDP Growth



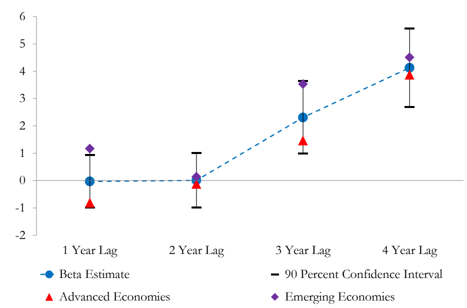
(b) Credit-to-GDP Gap



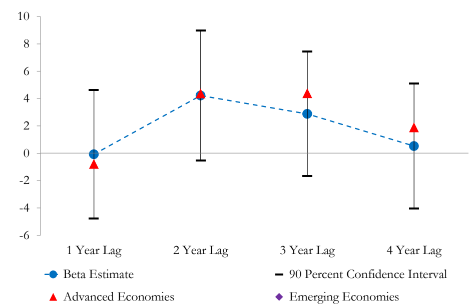
(c) Banking Sector Leverage



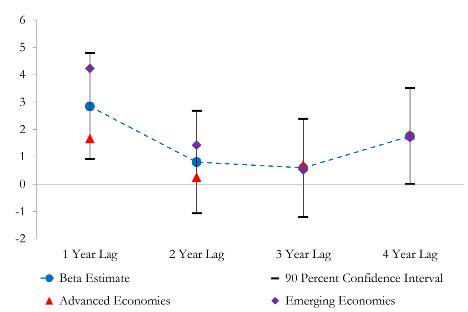
(d) Bank Interconnectedness Proxy I



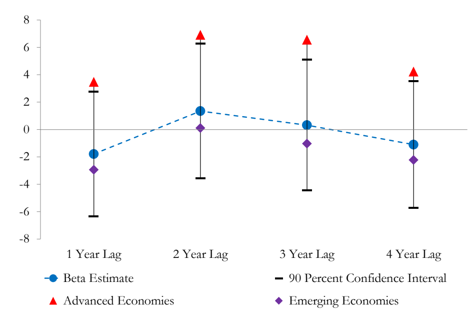
(e) Equity Price Growth



(f) House Price Growth



(g) REER Appreciation



(h) Real GDP Growth

Figure 2.4: *Systemic Risk Factors. All the estimations are based on a single factor logit model with country fixed effects. Models with different lags are estimated using the same data sample. The red triangles denote the parameter estimate when only using advanced economies in the estimation and the green squares denote the parameter estimate when only using emerging economies. See Table* 2.6 and 2.7 *in appendix* 2.7.2 *for estimation details.*

Table 2.2: Systemic Risk Factors based on Dynamic Logit Mode, 1970-2010

	Lag, h	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Credit-to-GDP Gap (in pct. points)	1	16.91***	11.25***	-	20.58***	6.94***	15.95***	18.78***	17.05***	8.35**	14.84*	7.44
		5.21	3.20		6.25	2.64	5.74	5.72	5.95	3.46	7.91	6.23
Credit-to-GDP Gap (in pct. points)	2	-	-	11.97***	-	-	-	-	-	-	-	-
				3.27								
Banking Sector Leverage (in pct.)	1	6.85***	0.64**	-	9.45***	1.33**	-	-	-	-	6.65*	6.23***
		2.19	0.31		2.96	0.55					3.76	2.03
Banking Sector Leverage (in pct.)	2	-	-	3.94**	-	-	-	-	-	-	-	-
				1.72								
Equity Price Growth (in pct.)	1	1.93**	0.26	-	2.23*	-	-	-	-	-	4.55	-
		0.90	0.70		1.18						3.01	
Equity Price Growth (in pct.)	2	-	-	2.44**	-	-	-	-	-	-	-	-
				1.02								
House Price Growth (in pct.)	1	-	-	-	3.40	-	-	-	-	-	-	-
					6.46							
Contagion Effect [1)]	1	-	-	-	-	6.77***	-	-	-	-	-	-
						1.25						
Bank Interconnectedness Proxy 1 (in pct.)	1	-	-	-	-	-	4.27**	-	-	-	-	-
							2.02					
Bank Interconnectedness Proxy 2 (in pct.)	1	-	-	-	-	-	-	3.66***	3.47**	-	-	-
								1.25	1.37			
REER Appreciation (in pct.)	1	-	-	-	-	-	2.10*	-	6.02*	-	-	-
							1.26		3.52			
Bad Credit Premium [2)]	1	-	-	-	-	-	-	-	-	-	-	16.34**
												7.16
ΔLending Premium (in pct. points)	1	-	-	-	-	-	-	-	-	-17.63**	-65.00	-
										7.39	101.0	
Country Fixed Effects		Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
McFadden's $R2$ (in%)		32.8	9.9	28.9	33.0	25.7	33.4	31.8	36.7	16.8	30.2	31.8
Unrestricted likelihood		-62.8	-84.2	-66.7	-50.2	-143.1	-45.5	-46.4	-40.5	-73.3	-23.6	-63.7
Countries / Observations		38 / 399	38 / 399	38 / 399	29 / 285	66 / 940	29 / 940	22 / 299	22 / 278	33 / 525	23 / 206	38 / 399

Notes: The model parameters are estimated by a partial likelihood estimator based on an unbalanced annual panel for the period 1970-2010. ***, ** and * indicate statistically significant parameters at a 1%, 5% and 10% significance level respectively (against a double sided alternative). Standard errors are written below the parameter estimates. The forecast horizon (in years) is denoted by h. Coefficient estimates and standard errors are in%. [1)] The contagion variable for country i, at time t+1, is defined as $\sum j \neq i \omega j,tyj,t$ where yj,t denote the binary systemic banking crisis for country j at time t and ωj,t is country j's market capitalization, at time t, as a%age of the world's market capitalization. [2)] Bad Credit Premium is defined as "Credit-to-GDP Gap"*"1{Previous two years equity inflation>20pct.}" where 1{.} denote an indicator function which takes the value unity if the condition is true and zero otherwise.

percent level as seen in columns one and three. This is consistent with the theory in Allen & Gale (2000*a*) which finds that asset-price inflation that is fuelled by credit expansion can lead to systemic banking crises.

2. Column two illustrates the effect of the fixed effect estimation methodology. Other studies on crisis prediction usually employ a random effect estimator which is only consistent if the unobserved country fixed effects are independent of the risk indicators. The empirical results indicate that the independence assumption might be violated since the parameter estimates are substantially different for the random effects estimator in column two. The potential inconsistency of the random effect estimator could reflect an omitted variable bias that arises due to the dependence between the country fixed effects and the explanatory variables. For example, suppose that economies with a small country fixed effect, maybe due to well-developed financial markets, are more likely to have a leveraged banking sector on average. In other words, suppose that there is an inverse relationship between the country fixed effects and the banking sector leverage risk factor. This will imply that the random effect estimator of the banking sector leverage coefficient will be inconsistent and have a negative asymptotic bias. In order to test this formally, one could conduct a Hausman-type test (Hausman 1978) to evaluate whether the difference between the two parameter estimates, $\hat{\boldsymbol{\theta}}_{FE} - \hat{\boldsymbol{\theta}}_{RE}$, is significantly different from zero.

3. House-price inflation does not appear to have a large impact on systemic risk as seen in column four. In a single predictor model, house-price inflation was only weakly related to systemic banking crisis and when combined with banking sector leverage, credit-to-GDP gap and equity price inflation, the impact is even weaker. However, it should be noted that this model specification is based on information from only 29 countries due to restrictions on data availability.

4. There is a significant intertemporal contagion effect between the economies. Column five illustrates the parameter estimates for a model specification with credit-to-GDP gap, banking sector leverage and the contagion variable defined in the previous section.

The estimated coefficient of the contagion variable is positive and significant, thus indicating that if an economy with a large financial sector is experiencing a systemic banking crisis then the systemic risk forecast increases in other economies in the next period. Kaminsky & Reinhart (2000) have found similar results.

5. Financial institutions' interconnectedness, as approximated by non-core to core bank liabilities, also appears to have a positive significant impact on systemic risk, as seen in columns six, seven and eight. The effect is positive and significant for both definitions of *non-core bank liabilities*. These empirical findings are consistent with the theoretical model in Hahm, Shin & Shin (2011).

6. Changes in the lending premium also appear to impact the level of systemic risk. Column nine illustrates the parameter estimates for a model specification with credit-to-GDP gap and changes in the lending premium. The coefficient on the lending premium is negative and significant, thus indicating that if the lending premium decreases in an economy then the level of systemic risk increases. However, when controlling for banking sector leverage and equity price growth, the effect is no longer significant as seen in column ten.

7. Deteriorating trade competitiveness, as measured by a real effective exchange rate appreciation, appears to increase the level of systemic risk. This is illustrated in columns six and eight.

8. An increase in the credit-to-GDP gap increases systemic risk substantially more when it is accompanied by high equity-price growth. The model specification in column eleven allows the effect of the credit-to-GDP gap to differ by adding a premium if the previous two-years' equity price inflation has exceeded 20 percent. The estimated coefficient of 'good' credit growth is 7.44 percent while the estimated coefficient of 'bad' credit growth is 23.78 percent (7.44 percent + 16.34 percent). Hence, the empirical results suggest that the increase in systemic risk, following an increase in the credit-to-GDP gap, is approximately three times larger if the last two years' equity price inflation has exceeded 20 percent. Consequently, the evolution in equity prices

might be useful for identifying a 'healthy' credit expansion.

9. Finally, it should be noted that the results are not sensitive to whether one uses the credit-to-GDP gap or the credit-to-GDP growth measurements. This is illustrated in table 2.8 and 2.9 in appendix 2.7.2. This is a useful result since it is computationally easier, and requires less data, to compute the credit-to-GDP growth rather than the credit-to-GDP gap based on a recursive Hodrick-Prescott filter.

Based on the estimation results in this section we find strong evidence that the level of systemic risk contains a predictable time-varying component. Note that the systemic risk estimates at time t are based solely on information available at time $t - h$, $t \geq 1$. The next section evaluates how to use this information to monitor systemic risk in real time.

2.5 Monitoring Systemic Risk

The empirical analysis has revealed that banking sector leverage, the credit-to-GDP gap, equity price inflation, real effective exchange rate appreciation and the ratio of non-core to core bank liabilities, all have a significant impact on the level of systemic risk in an economy. Furthermore, as all the model specifications provide an estimate of the level of systemic risk in the future, based on information today, the binary response model approach is a potentially useful tool for policy makers. The purpose of this section is to construct crisis signals, based on the binary response model approach, to evaluate the models' ability to monitor systemic risk. Since the focus is on monitoring systemic risk, rather than on how to implement macroprudential policy, we will not take a stand on which macroprudential tools to employ; instead, we simply assume that the policy decision is binary: the policymaker acts or he does not. This assumption also allows us to compare the binary response model approach with the signal extraction approach.

Table 2.3: Signal Classification

Events	No systemic banking crisis occurs ($y_{i,t} = 0$)	Systemic banking crisis actually occurs ($y_{i,t} = 1$)
The model does not issue a warning signal ($S_{i,t} = 0$)	A	B (Type I error)
The model issues a warning signal ($S_{i,t} = 1$)	C (Type II error)	D

2.5.1 The Signal Extraction Approach

The signal extraction approach was originally developed to identify turning points in business cycles and was first applied to banking crises by Kaminsky & Reinhart (1999). More recent studies include Borio & Lowe (2002) and Borio & Drehmann (2009). The methodology is straightforward:for each period a binary signal, $S_{i,t}$, is computed. The signal takes either the value 1 (is 'on'), or 0 (is 'off'). The signal, is 'on' if one or several risk indicators cross a certain threshold and 'off' otherwise. For the one-dimensional case with a single risk factor, $x_{i,t-h}$, the signal extraction approach can simply be represented as

$$S_{i,t}(\tau|\, x_{i,t-h}) = \begin{cases} 0 \; if \; x_{i,t-h} \leq \tau \\ 1 \; if \; x_{i,t-h} > \tau \end{cases},$$

where τ denotes the crisis threshold. A natural question then arises: how do we most optimally determine the threshold, τ^*? A popular assessment methodology distinguishes between two types of forecast errors: a type I error, when no signal is issued and a crisis occurs (i.e. $y_{i,t} = 1$ and $S_{i,t} = 0$), and a type II error, when a signal is issued but no crisis occurs (i.e. $y_{i,t} = 0$ and $S_{i,t} = 1$). Once a crisis occurs it makes no sense to predict another crisis: the indicator has already done its job. Therefore, we do not consider any signals in the two years after the beginning of a crisis. The different signal classifications are illustrated in table 2.3.

The optimal thresholds are determined by minimizing a loss function, $\mathcal{L}(\cdot)$, defined over the fraction of type I and type II errors, $\mathcal{L} : [0:1]^2 \to \mathbb{R}$,

$$\boldsymbol{\tau}^* = \underset{\boldsymbol{\tau}\in\mathbb{R}^k}{arg\,min}\ \left\{\mathcal{L}\left(\hat{e}^I(\boldsymbol{\tau}|\mathbf{y},\mathbf{X}), \hat{e}^{II}(\boldsymbol{\tau}|\mathbf{y},\mathbf{X})\right)\right\},$$

where $\boldsymbol{\tau} = (\tau_1, ..., \tau_k)^T$ denote the thresholds and

$$\hat{e}^I(\boldsymbol{\tau}|\mathbf{y},\mathbf{X}) = \frac{\sum_{i=1}^N \sum_{t=1}^T \mathbf{1}\left\{y_{i,t} = 1 \wedge S_{i,t}(\mathbf{t};\, \mathbf{x}_{i,t-h}) = 0\right\}}{\sum_{i=1}^N \sum_{t=1}^T \mathbf{1}\left\{y_{i,t} = 1\right\}}$$

and

$$\hat{e}^{II}(\boldsymbol{\tau}|\mathbf{y},\mathbf{X}) = \frac{\sum_{i=1}^N \sum_{t=1}^T \mathbf{1}\left\{y_{i,t} = 0 \wedge S_{i,t}(\mathbf{t};\, \mathbf{x}_{i,t-h}) = 1\right\}}{\sum_{i=1}^N \sum_{t=1}^T \mathbf{1}\left\{y_{i,t} = 0\right\}}$$

denote the fraction of type I and type II errors respectively. A common choice of loss function, e.g. Borio & Lowe (2002), is the noise-to-signal ratio:

$$\mathcal{L}^{\mathrm{NTS}}(\hat{\mathrm{e}}^{\mathrm{I}}, \hat{\mathrm{e}}^{\mathrm{II}}) = \hat{e}^{II}(\mathbf{t};\mathbf{y},\mathbf{X})/(1 - \hat{e}^I(\mathbf{t};\mathbf{y},\mathbf{X})).$$

2.5.2 Crisis Signals Based on the Binary Response Model

How does one construct crisis signals based on the binary response model framework? In this paper, we suggest that a crisis signal is 'on' if the estimated level of systemic risk, not individual risk indicators, cross a certain threshold, λ. More formally, this approach can be described as

$$S_{i,t}(\lambda|\mathcal{F}_{t-1};\hat{\boldsymbol{\beta}},\hat{\alpha}_i) = \begin{cases} 0\ if\ Pr(y_{i,t} = 1|\mathcal{F}_{t-1};\hat{\boldsymbol{\beta}},\hat{\alpha}_i) \le \lambda \\ 1\ if\ Pr(y_{i,t} = 1|\mathcal{F}_{t-1};\hat{\boldsymbol{\beta}},\hat{\alpha}_i) > \lambda \end{cases},$$

where $Pr(y_{i,t} = 1|\mathcal{F}_{t-1};\hat{\boldsymbol{\beta}},\hat{\alpha}_i) = G(\hat{\alpha}_i + x_{i,t-h}^T\hat{\boldsymbol{\beta}})$. The optimal threshold level, λ^*, depends again on the policy maker's preferences over these two types of errors as represented by a loss function, $\mathcal{L}(\cdot)$:

$$\lambda^* = \underset{\lambda\in[0:1]}{arg\,min}\ \left\{\mathcal{L}\left(\hat{e}^I(\lambda|\mathbf{y},\mathbf{X};\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}}), \hat{e}^{II}(\lambda|\mathbf{y},\mathbf{X};\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}})\right)\right\}, \tag{2.5.1}$$

where

$$\hat{e}^I(\lambda|\mathbf{y},\mathbf{X};\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}}) = \frac{\sum_{i=1}^N \sum_{t=1}^T \mathbf{1}\left\{y_{i,t} = 1 \wedge S_{i,t}(\lambda|\mathcal{F}_{t-h};\hat{\boldsymbol{\beta}},\hat{\alpha}_i) = 0\right\}}{\sum_{i=1}^N \sum_{t=1}^T \mathbf{1}\left\{y_{i,t} = 1\right\}}$$

and

$$\hat{e}^{II}(\lambda|\mathbf{y},\mathbf{X};\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}}) = \frac{\sum_{i=1}^N \sum_{t=1}^T \mathbf{1}\left\{y_{i,t} = 0 \wedge S_{i,t}(\lambda|\mathcal{F}_{t-h};\hat{\boldsymbol{\beta}},\hat{\alpha}_i) = 1\right\}}{\sum_{i=1}^N \sum_{t=1}^T \mathbf{1}\left\{y_{i,t} = 0\right\}}.$$

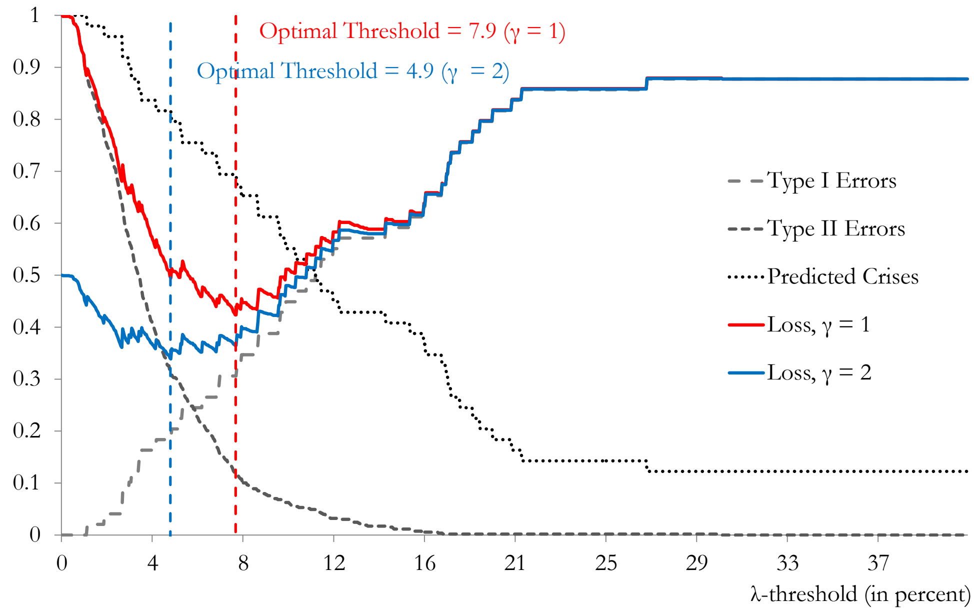


Figure 2.5: *Optimal Threshold. The figure illustrates the fraction of type I and type II errors as a function of the systemic risk threshold (based on a binary response model with banking sector leverage and the credit-to-GDP gap). The estimated conditional crisis probability for a given year is formed in the preceding year. Type I Errors = B/(B+D), Type II Errors = C/(A+C), Predicted Crises = D/(B+D) and Loss Function = γ*Type I Errors + Type II Errors, where γ={1,2}.*

When calibrating both τ^* and λ^* we will assume that the policy maker's preferences can be approximated by $\mathcal{L}(\hat{\mathrm{e}}^{\mathrm{I}}, \hat{\mathrm{e}}^{\mathrm{II}}) = \hat{e}^{I} + \hat{e}^{II}$. [12]

Naturally, when choosing the optimal threshold there is a trade-off between type I and type II errors. This is illustrated in Figure 2.5 for a binary response model with banking sector leverage and the credit-to-GDP gap.

The trade-off is clear: when a variable captures a lot of crises (few type I errors) due to a low threshold, it tends to overpredict their number. That is, it tends to issue false signals and exhibit a high number of type II errors. Figure 2.5 also illustrates that the optimal threshold depends on the policy maker's preferences. The optimal threshold is 7.9 if policy maker's

[12]The noise-to-signal ratio, Type I / (1 – Type II), is often used to approximate the policy makers preferences (e.g. Borio and Lowe, 2002). We used the alternative specification, Type I + Type II, as we found that the noise-to-signal ratio led to an optimal calibration where few crises were correctly predicted. Drehmann, Borio and Tsatsaronis (2011) experienced a similar problem and they choose to minimize the noise-to-signal-ratio subject to at least two-thirds of the crises being correctly predicted.

preferences can be approximated by the loss function $\mathcal{L}(\hat{\mathrm{e}}^{\mathrm{I}}, \hat{\mathrm{e}}^{\mathrm{II}}) = \hat{e}^{I} + \hat{e}^{II}$, and 4.9 if the loss function can be approximated by $\mathcal{L}(\hat{\mathrm{e}}^{\mathrm{I}}, \hat{\mathrm{e}}^{\mathrm{II}}) = 2\hat{e}^{I} + \hat{e}^{II}$.

One might ask whether the binary response model approach has any advantages relative to the signal extraction approach. If the binary response model provides a good approximation of the true conditional crisis probability this approach should lead to more accurate crisis probability forecasts. In addition, the thresholds in the binary response model are dynamic in the sense that the threshold of a single risk-factor depends on the values of the other risk factors. This allows for a more realistic environment where appropriate policy thresholds depend on the state of the economy via several risk factors.

Table 2.4 illustrates the signalling performance of different model specifications based on both the signal extraction method and the binary response model approach. Interestingly, the binary response model approach outperforms the signal extraction approach for all model specifications. For example, based on a model specification with credit-to-GDP growth, banking sector leverage and equity price growth the policy maker's estimated expected 'loss' is 29.9 percent, whereas the signal extraction approach based on Credit-to-GDP gap is 63.1 percent and 47.0 percent when using the same data sample.

For the model specification with credit-to-GDP growth, banking sector leverage and equity price growth, the optimal systemic risk threshold is determined to be around 13 percent. How does this relate to the threshold levels for the underlying risk factors? We will explore this question in the following section.

2.5.3 Risk Factor Thresholds

In the previous section we identified several systemic risk factors that affect the conditional crisis probability and we have shown that they provide accurate crisis signals in terms of type I and type II errors. A crisis signal is issued when the estimated systemic risk reaches a certain threshold. This section looks at how the optimal threshold is related to the levels of the underlying systemic risk factors. We focus on model specification one from table 2.8,

Table 2.4: Monitoring Systemic Risk

	Threshold (λ*, t*)	Type I	Type II	Prediction	NTS ratio	Loss	Countries	Crises
Binary Response Model (logit):								
Credit-to-GDP Gap	9.2	50.6%	7.8%	49.4%	15.7%	58.4%	68	87
Credit-to-GDP Growth	6.6	37.9%	18.2%	62.1%	29.3%	56.1%	68	87
Credit-to-GDP Gap and Bank Leverage	7.9	30.6%	11.7%	69.4%	16.9%	42.4%	66	49
Credit-to-GDP Gap and Bank Interconnectedness Proxy I	8.1	31.6%	5.8%	68.4%	8.5%	37.4%	22	19
Credit-to-GDP Growth and REER Appreciation	9.9	44.3%	10.7%	55.7%	19.2%	55.0%	66	70
CrtG Gap, Bank Leverage and Equity Price Growth	12.3	20.0%	8.8%	80.0%	10.9%	28.8%	38	25
CtG Growth, Bank Leverage and Equity Price Growth	13.0	24.0%	5.9%	76.0%	7.7%	29.9%	38	25
CtG Growth, Bank Leverage, Equity- and House Price Growth	9.5	14.3%	17.9%	85.7%	20.9%	32.2%	29	21
Signal Extraction Approach:								
Credit-to-GDP Growth	2.8	41.4%	21.7%	58.6%	37.0%	63.1%	68	87
Credit-to-GDP Growth [1)]	4.9	28.0%	19.0%	72.0%	26.3%	47.0%	38	25
Credit-to-GDP Gap	2.3	36.8%	25.7%	63.2%	40.6%	62.4%	68	87
Equity Price Growth	16.5	25.0%	35.6%	75.0%	47.5%	60.6%	40	44
House Price Growth	9.8	37.1%	28.8%	62.9%	45.8%	65.9%	35	30
Banking Sector Leverage	128.3	45.1%	27.7%	54.9%	50.4%	72.8%	67	51
Bank Interconnectedness Proxy I	31.4	31.6%	43.1%	68.4%	62.9%	74.7%	22	19

Notes: The parameters in the binary response models are estimated by a partial likelihood approach using country fixed effects. The calibration of the thresholds are based on the assumption that the policymaker's preferences over type I and type II errors can be approximated by the following loss function: L(TypeI, TypeII) = TypeI + TypeII. Loss denotes the value of the loss function (in percent) for the optimal calibration. The thresholds for the signal extraction approach are risk indicator thresholds while thresholds for the binary response models refer to the level of systemic risk. NTS denotes the noise-to-signal ratio and is given by Type II/(1-Type I). [1)] For comparison, we also compute the optimal threshold based on the data sample for the model specification with credit-to-GDP growth, banking sector leverage and equity price growth.

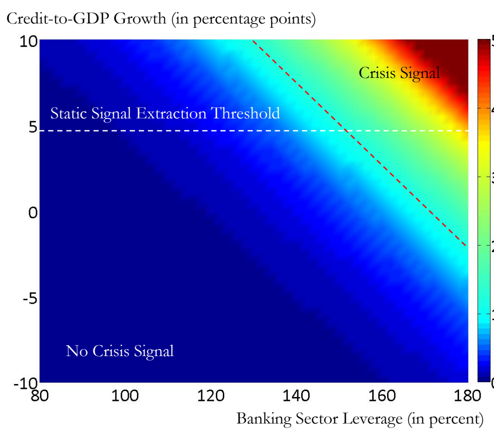
with credit-to-GDP growth, banking sector leverage and equity price growth, due to its good signalling properties as illustrated in table 2.4.

When should a policy maker be concerned about the level of systemic risk? If his/her preferences over forecasting errors can be described by, $\mathcal{L}(TypeI(\lambda), TypeII(\lambda)) = TypeI(\lambda) + TypeII(\lambda)$, he/she should react when the conditional crisis probability reaches around 13 percent. Figure 2.6 illustrates the systemic risk estimates for different values of the risk indicators and the corresponding crisis signal regions. Based on this figure it is clear that all the risk indicators affect each other such that the threshold of one indicator depends on the level of the other indicators. For example, if equity prices have grown by 20 percent and the banking sector leverage is around 160 percent, a crisis signal is issued if the credit-to-GDP growth is above 0 percent (figure 2.6.b). On the other hand, if equity prices have decreased by 10 percent and the banking sector leverage is around 130 percent a crisis signal is only issued if the credit-to-GDP gap is above 10 percent points (figure 2.6.a).

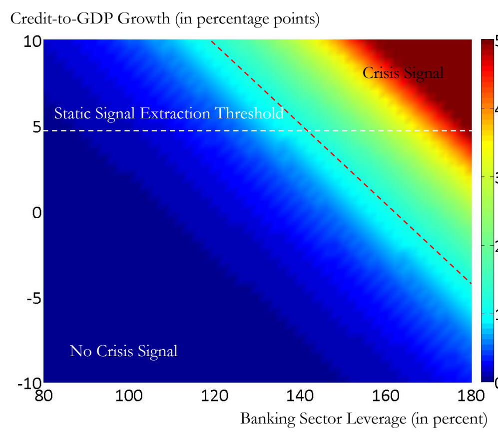
Figure 2.6 also illustrates the crisis signal region for a signal extraction model with credit-to-GDP growth. A crisis signal is simply issued when the growth rate is above 4.9 percentage points.[13]

The analyses in tables 2.2 and 2.4 indicate that a policy maker might gain from combining several systemic risk indicators. Rather than just looking at the credit-to-GDP growth, for example, it might also be useful to examine the amount of leverage in the banking sector. A crisis signal rule based solely on whether or not the credit-to-GDP gap is above 4.9% might be too simplistic. A credit boom does not necessarily increase systemic risk if it reflects a healthy market response to expected future productivity gains as a result of new technology, new resources or institutional improvements. Indeed, many episodes of credit booms were not followed by a systemic banking crisis or any other material instability. The binary

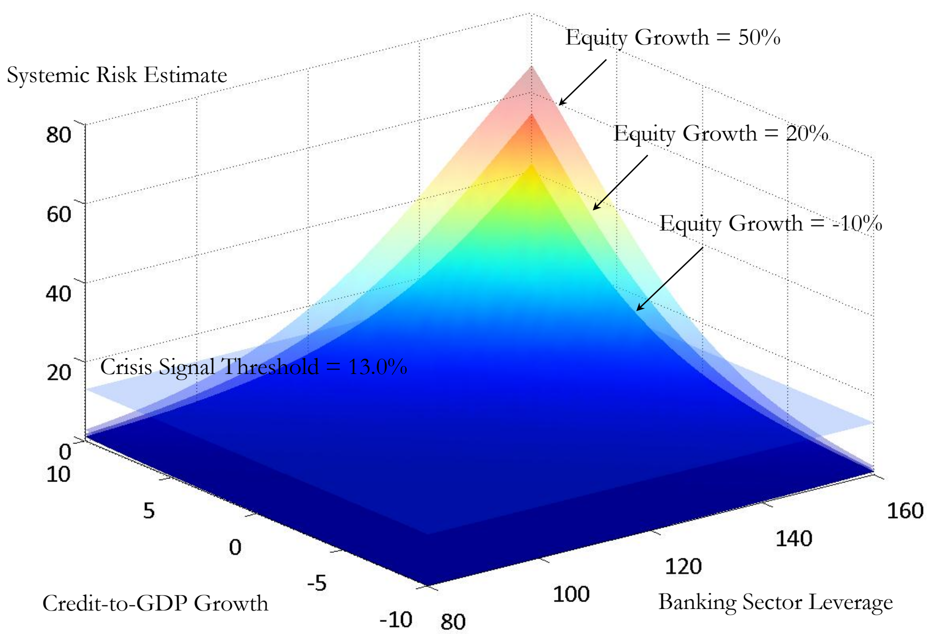
[13] In Drehmann, Borio and Tsatsaronis (2011), the optimal threshold level for credit-to-GDP growth is found to be around 8%. However, they use credit-to-GDP growth in% rather than in%age points. In addition, their optimal threshold is based on minimizing the noise to signal ratio, subject to at least two-thirds of the crises being correctly predicted, rather than $\mathcal{L}$(TypeI(λ),TypeII(λ))=TypeI(λ)+TypeII(λ).



(a) Equity growth is -10%



(b) Equity growth is 20%



(c) Three dimensional illustration

Figure 2.6: *Systemic Risk Estimates and Crisis Signals. The dotted red line separates the 'crisis signal' and 'no crisis signal' regions based on the binary response model approach. The estimates are based on a fixed effect logit model with banking sector leverage, the credit-to-GDP growth and equity price growth. The probabilities are evaluated at the median country fixed effect. For more details on estimation results see column (1) in 2.8 in appendix* 2.7.2.. *The dotted white line illustrates the static threshold based on the signal extraction approach.*

response model framework allows for a more realistic environment where the risk indicator threshold depends on the level of the other risk indicators. In general, the optimal threshold for risk factor j, is given by

$$Risk - Factor^*_{i,t-h,j} = \frac{1}{\hat{\beta}_j}\{\Lambda^{-1}(\lambda^*) - \hat{\alpha}_{Med} - \sum_{k \neq j} \hat{\beta}_k x_{i,t-h,k}\},$$

where $x_{i,t-h,k}$ denotes the value of risk factor k for country i at time $t-h$, $\Lambda^{-1}(x) = log(\frac{x}{1-x})$ and λ^*denotes the optimal systemic risk threshold from equation 2.5.1. For model specification one from table 2.8, with credit-to-GDP growth, banking sector leverage and equity price growth, the optimal systemic risk threshold is 13.0% (see table 2.4). The corresponding optimal credit-to-GDP growth threshold is given by the following equation:

$$\begin{aligned} CtG - Growth^* &= \frac{1}{22.91\%}(\Lambda^{-1}(13.0\%) - (-10.96) - 5.32\% \times' Bank\,Leverage' \\ &\quad -1.70\% \times' Equity\,Growth') \\ &= 39.5 - 23.2\% \times' Bank\,Leverage' - 7.4\% \times' Equity\,Growth', \end{aligned}$$

where $CtG - Growth^*$ denotes the optimal credit-to-GDP growth threshold in percentage points and -10.96 denotes the median country fixed effect. Figure 2.7 illustrates the credit-to-GDP growth thresholds for different values of the other risk indicators.

Based on figures 2.6 and 2.7, it is clear that if equity price growth has increased dramatically then the optimal credit-to-GDP threshold decreases consequently. This dynamic is consistent with the theoretical model in Allen & Gale (2000*a*) where asset price inflation and credit growth amplify each other.

2.5.4 Out-of-Sample Analysis

As a robustness check we also evaluate the models' performance to provide early warning signals, out-of-sample. Specifically, we use data from 1970-2000 to estimate the model parameters and use these to construct early warning signals, out-of-sample, for the period 2001-2010. The optimal thresholds are calibrated based on data from 1970-2000. Table 2.5

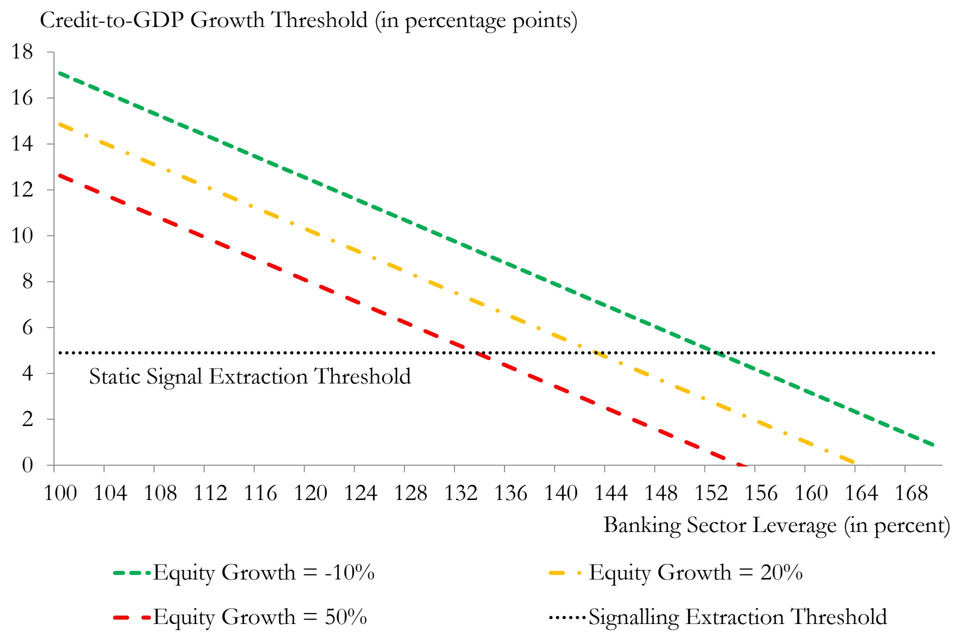


Figure 2.7: *Dynamic Credit-to-GDP Growth Threshold. The black dotted line illustrates the static threshold based on the signal extraction approach. The green, yellow and red lines illustrates the dynamic data dependent thresholds based on the binary response model framework. The thresholds are based on a model with banking sector leverage, equity price growth and credit-to-GDP growth evaluated at the median country fixed effect. The threshold calibration is based on the assumption that the policy maker's preferences can be approximated by* $\mathcal{L}(\hat{\mathrm{e}}^{\mathrm{I}}, \hat{\mathrm{e}}^{\mathrm{II}}) = \hat{e}^{I} + \hat{e}^{II}$.

contains the results for various logit model specifications.

Not surprisingly, the policy maker's loss is generally larger than in the in-sample analysis. That being said, however, the model specifications still perform well. All the models predict around 50-80 percent of the systemic banking crises, out-of-sample, without issuing an unreasonably high level of false signals. Figure 2.8 illustrates the estimated systemic risk for the United States based on a binary response model with credit-to-GDP growth and banking sector leverage. Interestingly, the systemic risk estimates increase quite dramatically as the sub-prime mortgage crisis approaches in 2007.

Table 2.5: Monitoring Systemic Risk - Out-of-Sample Analysis, 2001-2010

	Threshold (λ*)	Type I	Type II	Prediction	NTS ratio	Loss	Countries	Out-of-sample Crises
Single Factor Models:								
Credit-to-GDP Gap	6.6	50.0%	36.9%	63.1%	79.2%	86.9%	43	8
Credit-to-GDP Growth	6.3	62.5%	36.9%	37.5%	98.3%	99.4%	43	8
Banking Sector leverage	6.9	50.0%	18.7%	50.0%	37.4%	68.7%	21	4
REER Appreciation	6.6	50.0%	48.2%	50.0%	96.4%	98.2%	66	26
Two Factor Models:								
Credit-to-GDP Gap and Bank Leverage	5.7	0.0%	34.8%	100.0%	34.8%	34.8%	21	4
Credit-to-GDP Gap and Equity Price Growth	4.8	25.0%	37.1%	75.0%	49.5%	62.1%	20	4
Credit-to-GDP Growth and Bank Leverage	7.5	25.0%	18.1%	75.0%	24.1%	43.1%	21	4
Credit-to-GDP Growth and REER Appreciation	8.0	62.5%	33.4%	37.5%	89.1%	95.9%	40	8
Credit-to-GDP Growth and House Price Inflation	4.0	0.0%	66.7%	100.0%	66.7%	66.7%	11	3
Three Factor Models:								
Credit-to-GDP Growth, Bank Leverage and Equity Price Growth	9.8	50.0%	36.9%	50.0%	73.8%	86.9%	9	4
Credit-to-GDP Gap, Bank Leverage and House Price Growth	5.7	33.3%	55.3%	66.7%	83.0%	88.7%	7	3

Notes: The out-of-sample crisis signals for 2001-2010 are based on a dynamic logit model, with country fixed effects, estimated over the period 1970-2000. The parameters are estimated by maximum likelihood. A type I error is when a systemic banking crisis occurs but no crisis signal was issued (in the current or two previous periods) and a type II error is when a signal is issued but no crisis occurs (in the current or the next two periods). Once a crisis occurs it makes no sense to predict another crisis since the indicator already has done its job. Therefore, any signals in the two years after the beginning of a crisis are ignored. The calibration of the thresholds are based on the assumption that the policymaker's preferences over type I and type II errors can be approximated by the following loss function: L(TypeI, TypeII) = TypeI + TypeII. The calibration is based on the in-sample period, 1970-2010. NTS denote the noise-to-signal ratio and is given by Type II/(1-Type I).

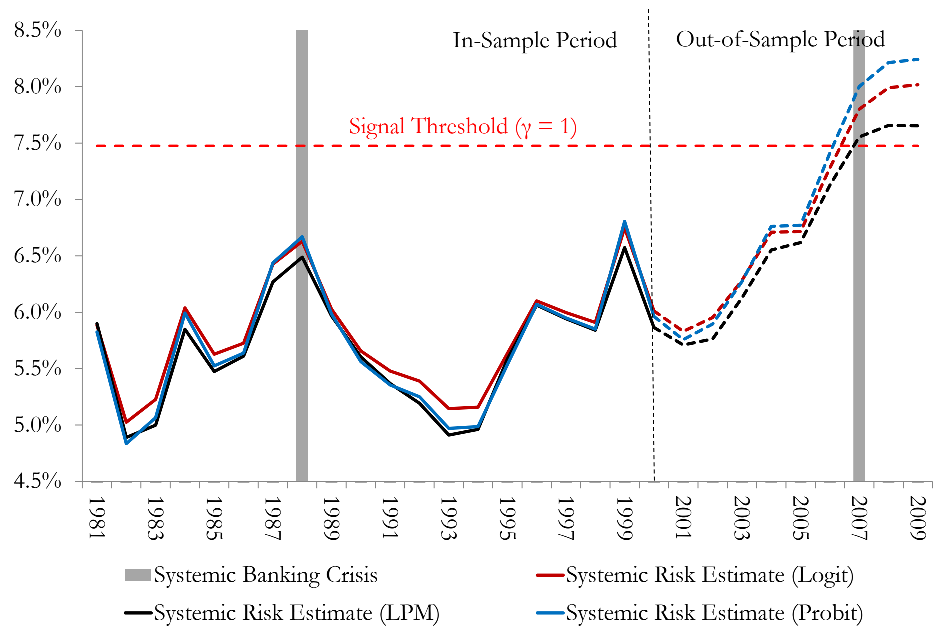


Figure 2.8: *Systemic Risk Estimates for the United States. The forecast of systemic risk for a given year is formed in the preceding year. Systemic risk estimates are based on a binary response panel data model with credit-to-GDP growth, banking sector leverage and country fixed effects for 1970-2000. The dashed lines show the out-of-sample probabilities for 2001-2009 and the horizontal red line depicts the optimal threshold level, 7.5, see also 2.5.*

2.6 Concluding Remarks

In this paper we used a dynamic binary response model with country fixed effects to model the time varying conditional probability of a systemic banking crisis. We provided conditions for consistency and asymptotic normality of the model parameters in the presence of cross-sectional dependence (e.g. due to contemporaneous contagion effects). We found that the level of systemic risk depends significantly on several risk factors: banking sector leverage, credit-to-GDP growth, changes in banks' lending premium, equity price growth, increasing interconnectedness in the financial sector and real effective exchange-rate appreciation. We have shown how to translate the systemic risk estimates into crisis signals and that this method provided accurate crisis signals in terms of type I and type II errors.

We discussed the implications for economic policy and threshold determination. In the signal extraction method the thresholds are constant: a crisis signal is issued if a variable is above a specific threshold, independent of the value of other systemic risk indicators. This is problematic since a credit boom does not necessarily increase the level of systemic risk. For example, the credit-to-GDP ratio could increase as a consequence of a healthy market response to expected future productivity gains as a result of new technology, new resources or institutional improvements. The thresholds, based on the binary response model framework, allow for a more realistic environment where the appropriate threshold depends on the state of the economy as measured by a combination of other risk factors.

Finally, a word of warning is warranted. This paper has shown that there exist several risk factors that can help forecast the level of systemic risk in an economy. Although the choices of risk factors generally were motivated by the theoretical literature, it is still a bit unclear whether they represent a causal relationship or not. In particular, more theoretical research is needed in order to evaluate whether credit-to-GDP growth has a causal impact on the level of systemic risk or if it is simply related to another economic variable that does. For monitoring purposes it does not matter whether credit-to-GDP growth has a causal impact or not. However, macroprudential policy that attempts to reduce credit-to-GDP growth will only be successful if there is a causal relationship. It is beneficial to consider a simple analogy in order to illustrate this point. Suppose that two individuals are accused of having committed a crime and you want to identify the offender. Ideally, you would like to observe the individuals' propensity to commit crime. This information is unobservable, but you do observe that one of them have a tattoo on his arm. The tattoo is useful for predicting which one is the offender if having a tattoo is related to an individual's propensity to commit crime. However, there is clearly no causal relationship between having a tattoo and an individual's propensity to commit crime. Therefore, although the tattoo variable is useful for forecasting, a policy that prohibits tattoos will have no impact on the crime level. To reiterate, based on the empirical results in this paper, it is yet unclear whether macroprudential policy should aim to reduce, or mitigate, any of the identified risk factors used in this model. More research is needed in order to answer this question.

2.7 Appendices

2.7.1 Data Sources and Description

Binary Systemic Crisis Variable:

We adopt the definition of a systemic banking crisis from Reinhart & Rogoff (2011). They define a banking crisis to be systemic if two conditions are satisfied:

- Bank runs that lead to closure, merger, or takeover by the public sector of one or more financial institutions.
- Or if there are no runs, the closure, merger, takeover, or large-scale government assistance of an important financial institution that mark the start of a string of similar outcomes for other financial institutions

Only the first crisis year is used in the estimation. The exact dates can be found in appendix 2.7.3. All data are annual.

Systemic Risk Factors:

- Credit-to-GDP (in percent): 100*(Private credit by deposit money banks)/(GDP). Source: IMF's International Financial Statistics (line 22d and 99b).

$$Credit\,to\,GDP\,Growth = (Credit\,to\,GDP_t - Credit\,to\,GDP_{t-2}) \cdot \frac{1}{2}$$
$$Credit\,to\,GDP\,Gap = Credit\,to\,GDP_t - Trend_t^{CtG}$$

  where $Trend_t^{CtG}$ is computed by the Hodrick-Prescott Filter (1981) with smoothing parameter, $\lambda = 1600$.

- Banking Sector Leverage (in percent): 100*(Private credit by deposit money banks)/(demand, time and saving deposits in deposit money banks). Source: IMF's International Financial Statistics (line 22d, line 24 and 25).

- Equity Prices are from Bloomberg.

$$Equity\,Price\,Growth_t = log(Equity\,Price_t/Equity\,Price_{t-2}) \times 100/2$$

- House Prices are from Bloomberg.

$$House\,Price\,Growth_t = log(House\,Price_t/House\,Price_{t-2}) \times 100/2$$

- REER: Real Effective Exchange rate. Source: IMF Information Notice System (INS) database.

$$REER\,Growth_t = log(REER_t/REER_{t-1}) \times 100$$

- Non-core liabilities Proxy 1: The sum of banks foreign liabilities (line 26C) and bank liabilities to other financial institutions (line 36J). Source: IMF's International Financial Statistics.
- Non-core liabilities proxy 2: The sum of banks foreign liabilities (line 26C) and M3-M2. Source: IMF's International Financial Statistics.
- Core Liabilities: Broad money, M2. Source: IMF's World Economic Outlook.
- Real GDP (RGDP) are from IMF's World Economic Outlook.

$$Real\,GDP\,Growth_t = log(RGDP_t/RGDP_{t-2}) \times 100/2$$

- The contagion variable for country i is defined as

$$Contagion_{i,t+1} = \sum_{j \neq i} \omega_{j,t} y_{j,t}\ ,\ \omega_{j,t} = \frac{Market\,Capitalization_{j,t}}{\sum_{i=1}^{N} Market\,Capitalization_{i,t}}$$

  where $y_{j,t}$ is the binary systemic banking crisis variable for country j at time 't'. Source: World Bank.
- Change in Lending Premium: Risk premium on lending (the interest rate charged by banks on loans to prime private sector customers minus the "risk free" treasury bill interest rate at which short-term government securities are issued or traded in the market). Source: World Bank.

$$\Delta Lending\,Premium_t = (Lending\,Premium_t - Lending\,Premium_{t-2})/2$$

2.7.2 Systemic Risk Factors

Table 2.6: Systemic Risk Factors, 1970-2010 (a)

Dependent variable: Binary Systemic Banking Crisis Variable from Reinhart and Rogoff (2010)

	Logit					Probit					Linear Probability Model		
Lag	$\hat{\beta}$	SE	ME LR	ME HR	McFR^2	$\hat{\beta}$	SE	ME LR	ME HR	McF R^2	$\hat{\beta}$	SE	R^2
Credit-to-GDP Growth (year-on-year, in pct. points)													
1	11.13***	2.71	0.32	0.91	11.89	5.62***	1.36	0.38	0.93	11.82	0.52***	0.10	7.27
2	8.25***	2.47	0.24	0.80	10.44	4.42***	1.28	0.30	0.84	10.55	0.39***	0.10	6.47
3	1.70	2.03	0.05	0.20	8.41	0.78	0.98	0.05	0.18	8.42	0.12	0.12	5.51
4	0.16	2.00	0.00	0.02	8.30	0.21	1.02	0.01	0.05	8.32	0.01	0.12	5.44
Countries: 64		Observations = 1442											
Credit-to-GDP Gap (in pct. points), λ = 1600													
1	7.03***	1.59	0.21	0.78	12.17	3.44***	0.79	0.24	0.73	12.05	0.35***	0.07	7.25
2	5.36***	1.52	0.16	0.66	10.61	2.78***	0.77	0.19	0.62	10.70	0.27***	0.07	6.49
3	2.83**	1.40	0.08	0.37	8.97	1.47**	0.70	0.10	0.35	9.03	0.15**	0.07	5.74
4	1.55	1.41	0.05	0.21	8.50	0.84	0.71	0.06	0.21	8.54	0.08	0.08	5.53
Countries: 64		Observations = 1442											
Banking Sector Leverage (in pct.)													
1	1.47***	0.51	0.03	0.14	15.96	0.78***	0.27	0.04	0.15	16.05	0.1***	0.029	9.24
2	1.15**	0.50	0.03	0.12	14.96	0.6**	0.27	0.03	0.12	15.00	0.08***	0.029	8.62
3	0.50	0.53	0.01	0.06	13.83	0.26	0.28	0.02	0.05	13.87	0.03	0.030	8.00
4	0.55	0.52	0.02	0.06	13.89	0.31	0.28	0.02	0.06	13.97	0.03	0.030	8.03
Countries: 64		Observations = 933											
Bank Interconnectedness Proxy I (in pct.)													
1	2.04*	1.14	0.10	0.17	12.39	0.92	0.58	0.10	0.15	12.14	0.15**	0.074	4.62
2	0.92	1.22	0.05	0.08	10.82	0.41	0.62	0.05	0.07	10.82	0.060	0.074	3.71
3	0.05	1.34	0.00	0.00	10.48	-0.03	0.65	0.00	0.00	10.55	0.003	0.072	3.53
4	-0.45	1.42	-0.02	-0.04	10.55	-0.26	0.69	-0.03	-0.04	10.66	-0.022	0.070	3.56
Countries: 29		Observations = 382											

Notes: The model parameters are estimated by a partial likelihood appraoch based on a country fixed effect model specification. The estimation is based on an unbalanced annual panel for the period 1970-2010. Models with different lags are estimated using the same data sample. Beta refers to the risk factor parameter estimate, SE refers to the standard error, McF.s R2 refers to McFadden's R2, R2 refers to the coefficient of determination, ME LR refers to the Marginal effect for a low risk country (20th percentile country fixed effect) and ME HR refers to the Marginal effect for a high risk country (80th percetile country fixed effect). The marginal effects are evaluated at the median value of the risk factor. Lag refers to h in $Pr(y_{i,t} = 1) = G(\alpha_i + \beta^T x_{i,t-h})$. The Credit-to-GDP gap is based on the deviation from the HP-filter trend with smoothing parameter λ=1600.

Table 2.7: Systemic Risk Factors, 1970-2010 (b)

Dependent variable: Binary Systemic Banking Crisis Variable from Reinhart and Rogoff (2010)

	Logit					Probit					Linear Probability Model		
Lag	Beta	SE	ME LR	ME HR	McF R2	Beta	SE	ME LR	ME HR	McF R2	Beta	SE	R2
Equity Price Growth (in pct.)													
1	-0.03	0.58	0.00	0.00	6.56	0.02	0.28	0.00	0.00	6.57	-0.002	0.03	2.17
2	0.01	0.60	0.00	0.00	6.52	0.05	0.29	0.00	0.01	6.58	0.001	0.03	2.17
3	2.31***	0.81	0.07	0.15	9.34	1.19***	0.40	0.08	0.16	9.64	0.09***	0.03	3.10
4	4.13***	0.88	0.09	0.23	14.90	2.01***	0.43	0.11	0.25	14.95	0.14***	0.03	4.78
Countries: 40		Observations = 862											
House Price Growth (in pct.)													
1	-0.08	2.86	0.00	-0.01	6.63	0.12	1.36	0.01	0.02	6.64	-0.003	0.12	2.84
2	4.22	2.90	0.11	0.29	7.43	2.22	1.37	0.13	0.32	7.65	0.15	0.12	3.09
3	2.88	2.78	0.08	0.21	7.03	1.58	1.31	0.10	0.23	7.19	0.11	0.11	2.97
4	0.52	2.78	0.01	0.04	6.64	0.35	1.33	0.02	0.05	6.66	0.02	0.11	2.84
Countries: 29		Observations = 724											
Real GDP Growth (in pct.)													
1	-1.79	2.78	-0.05	-0.10	3.43	-0.7199	1.30	-0.05	-0.08	1.38	-0.09	0.14	1.31
2	1.35	3.00	0.04	0.07	3.41	0.6415	1.42	0.04	0.07	1.38	0.06	0.14	1.30
3	0.32	2.91	0.01	0.02	3.38	0.25	1.37	0.02	0.03	1.37	0.01	0.14	1.29
4	-1.11	2.82	-0.03	-0.06	3.40	-0.35	1.33	-0.02	-0.04	1.37	-0.05	0.14	1.30
Countries: 68		Observations = 2306											
Real Effective Exchange Rate Appreciation (in pct.)													
1	0.24*	0.137	0.00	0.01	4.6	0.12*	0.07	0.00	0.01	4.6	0.02*	0.01	1.60
2	0.21	0.139	0.00	0.01	4.7	0.11	0.07	0.00	0.01	4.7	0.013	0.01	1.62
3	0.16	0.144	0.00	0.00	4.5	0.08	0.07	0.00	0.00	4.5	0.009	0.01	1.55
4	0.17	0.142	0.00	0.00	4.5	0.08	0.07	0.00	0.01	4.5	0.010	0.01	1.56
Countries: 67		Observations = 1585											

Notes: The model parameters are estimated by maximum likelihood based on a country fixed effect model specification. The estimation is based on an unbalanced annual panel for the period 1970-2010. Models with different lags are estimated using the same data sample. Beta refers to the risk factor parameter estimate, SE refers to the standard error, McF.s R2 refers to McFadden's R2, R2 refers to the coefficient of determination, ME LR refers to the Marginal effect for a low risk country (20th%ile country fixed effect) and ME HR refers to the Marginal effect for a high risk country (80th percetile country fixed effect). The marginal effects are evaluated at the median value of the risk factor. Lag refers to 'h' in Pr(yi,t=1)=G(αi+β×xi,t-h).

Table 2.8: Estimation Results, with Credit-to-GDP Growth, 1970-2010 (a)

	Lag, h	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Credit-to-GDP Growth (in pct. points)	1	22.91***	15.63***	-	28.74***	21.47***	19.39***	15.75***
		7.79	3.69		9.19	8.27	6.60	5.96
Credit-to-GDP Growth (in pct. points)	2	-	-	21.26***	-	-	-	-
				7.61				
Banking Sector Leverage (in pct.)	1	5.32***	0.10	-	7.78***	-	-	-
		1.92	0.38		2.61			
Banking Sector Leverage (in pct.)	2	-	-	3.00*	-	-	-	-
				1.74				
Equity Price Growth (in pct.)	1	1.70*	0.41	-	1.86*	-	-	-
		0.87	0.70		1.09			
Equity Price Growth (in pct.)	2	-	-	2.52**	-	-	-	-
				1.04				
House Price Growth (in pct.)	1	-	-	-	2.86	-	-	-
					6.33			
Bank Interconnectedness Proxy 1 (in pct.)	1	-	-	-	-	3.72*	-	-
						1.99		
Bank Interconnectedness Proxy 2 (in pct.)	1	-	-	-	-	-	2.64***	2.52**
							0.99	1.12
REER Appreciation (in pct.)	1	-	-	-	-	2.15**	-	6.15*
						1.05		3.25
Country Fixed Effects		Yes	No	Yes	Yes	Yes	Yes	Yes
McFadden's R2 (in percent)		32.0%	13.1%	31.3%	32.5%	31.4%	27.2%	32.6%
Unrestricted likelihood		-63.5	-81.2	-64.2	-51.7	-46.8	-49.5	-43.1
Countries		38	38	38	29	29	22	22
Observations		399	399	399	285	355	299	278

Notes: The model parameters are estimated by a partial likelihood appraoch based on a country fixed effect model specification. The estimation is based on an unbalanced annual panel for the period 1970-2010. Models with different lags are estimated using the same data sample. Beta refers to the risk factor parameter estimate, SE refers to the standard error, McF.s R2 refers to McFadden's R2, R2 refers to the coefficient of determination, ME LR refers to the Marginal effect for a low risk country (20th percentile country fixed effect) and ME HR refers to the Marginal effect for a high risk country (80th percetile country fixed effect). The marginal effects are evaluated at the median value of the risk factor. Lag refers to h in $Pr(y_{i,t} = 1) = G(\alpha_i + \beta^T x_{i,t-h})$.

Table 2.9: Estimation Results, with Credit-to-GDP Growth, 1970-2010 (b)

	Lag, h	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Credit-to-GDP Growth (in pct. points)	1	12.51***	13.91**	24.18*	22.25**	14.74*	15.76***	-
		4.57	5.97	13.92	10.11	8.03	4.47	
Equity Price Growth (in pct.)	1	-	-	4.403	2.89**	-	-	2.85**
				3.03	1.41			1.43
Banking Sector Leverage (in pct.)	1	1.23**	-	5.92*	4.46*	4.95***	1.19**	4.98**
		0.56		3.39	2.40	1.81	0.57	2.54
Banking Sector Leverage (in pct.)	2	-	-	-	-	-	-	-
Credit-to-GDP Gap (in pct.)	1	-	-	-	-	-	-	12.56**
								5.57
Contagion Effect 1)	1	6.55***	-	-	-	-	-	-
		1.26						
Credit-to-GDP Gap (in pct. points)	1	-	-	-	-	-	-	-
Bad Credit Premium 2)	1	-	-	-	-	11.78*	-	-
						6.63		
ΔLending Premium (in pct. points)	1	-	-17.81**	-68.29	-	-	-	-
			7.43	94.9				
Real GDP Growth (in pct.)	1	-	-	-	4.14	-	-	6.33
					16.06			16.07
Real Interest Rate (in pct.)	1	-	-	-	10.86	-	-	11.37
					9.89			10.01
Country Fixed Effects		Yes	Yes	Yes	Yes	Yes	Yes	Yes
McFadden's R2 (in percent)		26.3%	16.9%	29.3%	24.1%	31.7%	19.8%	24.1
Unrestricted likelihood		-141.8	-73.3	-23.9	-44.0	-63.8	-154.29	-43.9
Countries		66	33	23	33	38	66	33
Observations		940	525	206	332	399	940	332

Notes: The model parameters are estimated by a partial likelihood estimator based on an unbalanced annual panel for the period 1970-2010. ***, ** and * indicate statistically significant parameters at a 1 percent, 5 percent and 10 percent significance level respectively (against a double sided alternative). Standard errors are written below the parameter estimates. The forecast horizon (in years) is denoted by h. Coefficient estimates and standard errors are in percent. 1) The contagion variable for country i, at time t+1, is defined as $\sum j{\neq}i\omega j,tyj,t$ where yj,t denote the binary systemic banking crisis for country j at time t and ωj,t is country j's market capitalization, at time t, as a percentage of the world's market capitalization. 2) Bad Credit Premium is defined as "Credit-to-GDP Growth"*"1{Previous two years equity inflation>20pct.}" where 1{.} denote an indicator function which takes the value unity if the condition is true and zero otherwise.

2.7.3 Systemic Banking Crises Dates

Table 2.10: Systemic Banking Crises Dates

Country	Systemic Banking Crises	Country	Systemic Banking Crises
Algeria	1990-1992	Kenya	1985-1989 and 1992-1995
Angola	1992-1998	Korea, Republic of	1983, 1985-1988 and 1997-2002
Argentina	1980-1982, 1989-1990, 1995-1996 and 2001-2003	Malaysia	1985-1988 and 1997-2001
Austria	1989-1992	Mexico	1981-1982 and 1994-2000
Australia	2008-2010	Morocco	1983-1984
Belgium	2008-2010	Myanmar	1996-2003
Bolivia	1986-1987, 1994-1997 and 1999	Netherlands	2008-2010
Brazil	1985, 1990, 1994-1997	New Zealand	1987-1990
Canada	1983-1985	Nicaragua	1987-1996 and 2000-2002
Central African Rep.	1976-1982, 1988-1999	Nigeria	1993-1995 and 1997
Chile	1976-1977 and 1981-1984	Norway	1987-1993
China	1992-1999	Panama	1988-1989
Columbia	1982-1987 and 1998-1999	Paraguay	1995-1999 and 2002
Costa Rica	1987 and 1994-1996	Peru	1983-1990 and 1999
Côte d'Ivoire	1988 -1991	Philippines	1981-1987 and 1997-2001
Denmark	1987-1992 and 2008-2010	Poland	1991-1995
Dominican Republic	1996 and 2003	Portugal	2008-2010
Equador	1981 and 1998-2002	Romania	1990-1999
Egypt	1981-1983 and 1990-1995	Russia	1995, 1998 and 2008-2009
El Salvador	1989	Singapore	1982-1983
Finland	1991-1994	South Africa	1977-1978 and 1989
France	1994-1995 and 2008-2010	Spain	1977-1985 and 2008-2010
Germany	1977-1979 and 2008-2010	Sri Lanka	1990-1994
Ghana	1982-1989 and 1997	Sweden	1991-1994
Greece	1991-1995 and 2008-2010	Switzerland	2008-2009
Guatemala	1990, 2001 and 2006	Thailand	1980-1987 and 1996-2001
Honduras	1999 and 2001-2002	Tunesia	1991-1995
Hungary	1991-1995 and 2008-2010	Turkey	1982-1985, 1991, 1994 and 2000
Iceland	1985-1986, 1993 and 2007-2010	United Kingdom	1974-1976, 1984, 1995 and 2007-2009
India	1993-1998	United States	1984-1991 and 2007-2010
Indonesia	1992, 1994 and 1997-2002	Uruguay	1971, 1981-1984 and 2002
Ireland	2007-2010	Venezuela	1978-1986 and 1993-1994
Italy	1990-1995	Zambia	1995
Japan	1992-2001	Zimbabwe	1995-2008

Notes: The definition of a systemic banking crisis follows Reinhart and Rogoff (2010). For further details see the original chapter. Only the first crisis year is used in the estimation with the exception of the US where, following Laeven & Valencia (2008, 2010), the crisis begins in 1988 rather than 1984.

2.7.4 Proof of Proposition 1 and Proposition 2.

Notation:

For any $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}) \in \Theta \subset B \times \mathcal{A}^N \subset \mathbb{R}^{k+N}$ we let $l_t(\boldsymbol{\theta}) = \sum_{i=1}^N log\, f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i)$,
$Q_0(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_{i=1}^N log\, f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i)\right]$ and
$\hat{Q}_T(\theta) = T^{-1}\sum_{t=1}^T \left(\sum_{i=1}^N log\, f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i)\right)$. For all i, the true conditional density function of $y_{i,t}$ is denoted by $p_i(y_{i,t}|\mathcal{F}_{t-1})$. Furthermore, let $a_{i,t}(\boldsymbol{\theta}_i) = (\alpha_i + \mathbf{x}_{i,t-h}^T \boldsymbol{\beta}_0)$,
$\tilde{q}_{i,t}(\boldsymbol{\theta}_i) = \left[\frac{[y-G(a_{i,t}(\boldsymbol{\theta}_i))]G''(a_{i,t}(\boldsymbol{\theta}_i))}{G(a_{i,t}(\boldsymbol{\theta}_i))(1-G(a_{i,t}(\boldsymbol{\theta}_i)))} - \left(\frac{[y-G(a_{i,t}(\boldsymbol{\theta}_i))]G'(a_{i,t}(\boldsymbol{\theta}_i))}{G(a_{i,t}(\boldsymbol{\theta}_i))(1-G(a_{i,t}(\boldsymbol{\theta}_i)))}\right)^2\right]$ and
$q_{i,t}(\boldsymbol{\theta}_i) = -\left(\frac{[y-G(a_{i,t}(\boldsymbol{\theta}_i))]G'(a_{i,t}(\boldsymbol{\theta}_i))}{G(a_{i,t}(\boldsymbol{\theta}_i))(1-G(a_{i,t}(\boldsymbol{\theta}_i)))}\right)^2$ where $G'(x) = \frac{\partial G(x)}{\partial x}$ and $G''(x) = \frac{\partial^2 G(x)}{\partial x \partial x}$.

Assumptions:

(**A.1**) For all i, the model $\{f(y_{it}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i);\, (\beta, \alpha_i) \in \mathcal{B} \times \mathcal{A},\, \mathcal{B} \times \mathcal{A} \subset \mathbb{R}^{k+1}\}$ is a corectly specified model of the true conditional density of $y_{i,t}$:

$$\exists(\beta_0, \alpha_{i,0}) \in \mathcal{B} \times \mathcal{A}\ such\, that\ f(y_{i,t}|\mathcal{F}_{t-1}; \beta_0, \alpha_{i,0}) = p_i(y_{i,t}|\mathcal{F}_{t-1})$$

Furthermore, for all i, the model $\{f(y_{it}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i);\, (\beta, \alpha_i) \in \mathcal{B} \times \mathcal{A},\, \mathcal{B} \times \mathcal{A} \subset R^{k+1}\}$ is dynamically complete such that $y_{i,t}|\mathcal{F}_{t-1} \sim i.i.d.\, Bernoulli\left(G(\alpha_i + \mathbf{x}_{i,t-h}^T \boldsymbol{\beta})\right)$ through time and $\int log\, f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i) dy_{i,t} = 1\ \forall(\boldsymbol{\beta}, \alpha_i) \in \mathcal{B} \times \mathcal{A}$.

(**A.2**) For all i, $Pr[f(y_{it}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i) \neq f(y_{it}|\mathcal{F}_{t-1}; \boldsymbol{\beta}_0, \alpha_{i,0})] > 0$ for all $(\beta, \alpha_i) \neq (\beta_0, \alpha_{i,0})$ in $\mathcal{B} \times \mathcal{A}$.

(**A.3**) $\Theta = \mathcal{B} \times \mathcal{A}^N$ is a compact set.

(**A.4**) For all i, $\mathbb{E}\left[log[f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i)]\right]$ is continuous in $(\boldsymbol{\beta}, \alpha_i)$ and
$\mathbb{E}\left[log[f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i)]\right] < 0\ \forall(\boldsymbol{\beta}, \alpha_i) \in \mathcal{B} \times \mathcal{A}$.

(**A.5**) $\hat{Q}_T(\boldsymbol{\theta})$ is stochastically equicontinuous on Θ:

$$\forall\epsilon > 0,\, \exists\delta > 0\ such\, that\ \lim_{T\to\infty} Pr(\sup_{\boldsymbol{\theta}\in\Theta}\ \sup_{\boldsymbol{\theta}'\in B(\theta,\delta)} |\hat{Q}_T(\boldsymbol{\theta}') - \hat{Q}_T(\boldsymbol{\theta})| > \epsilon) < \delta.$$

Proposition 1. *Let $\hat{\boldsymbol{\theta}}_{PL}$ be the partial likelihood estimator defined in* (2.3.1) *on page 60. If assumptions (A1)-(A5) hold then:* $\hat{\boldsymbol{\theta}}_{PL} \xrightarrow{p} \boldsymbol{\theta}_0$ *as* $T \rightarrow \infty$.

Proof. Firstly, we show that (A1)-(A2) ensures that $\boldsymbol{\theta}_0$ is globally indentified:

$Q_0(\boldsymbol{\theta}_0) > Q_0(\boldsymbol{\theta})\ \forall \boldsymbol{\theta} \in \Theta\, if\, \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. For $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}$ we have that

$$
\begin{aligned}
Q_0(\boldsymbol{\theta}_0) - Q_0(\boldsymbol{\theta}) &= E[l_t(\boldsymbol{\theta}_0) - l_t(\boldsymbol{\theta})] \\
&= \sum_{i=1}^{N} -E\left[log\left(\frac{log\, f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i)}{log\, f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}_0, \alpha_{i,0})}\right)\right]
\end{aligned}
$$

As $log(\cdot)$ is a stricly concave function, we have by Jensen's inequality that

$$
\begin{aligned}
Q_0(\boldsymbol{\theta}_0) - Q_0(\boldsymbol{\theta}) &> \sum_{i=1}^{N} -log\left(E\left[\frac{f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i)}{f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}_0, \alpha_{i,0})}\right]\right) \\
&= \sum_{i=1}^{N} -log\left(\int p_0(y_{i,t}|\mathcal{F}_{t-h})\left[\frac{log\, f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i)}{log\, f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}_0, \alpha_{i,0})}\right] dy_{i,t}\right).
\end{aligned}
$$

By (A1), we have that $f(y_{i,t}|\mathcal{F}_{t-1}; \beta_0, \alpha_{i,0}) = p_0(y_{i,t}|\mathcal{F}_{t-1})$ such that

$$
\begin{aligned}
\sum_{i=1}^{N} -log\left(\int p_0(y_{i,t}|\mathcal{F}_{t-h})\left[\frac{log\, f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i)}{log\, f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}_0, \alpha_{i,0})}\right] dy_{i,t}\right) &= \sum_{i=1}^{N} -log\left(\int log\, f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i) dy_{i,t}\right) \\
&= \sum_{i=1}^{N} 0 \\
&= 0
\end{aligned}
$$

where the second equality use that $f(y_{i,t}|\mathcal{F}_{t-1}; \boldsymbol{\beta}, \alpha_i)$ is a well defined density for all $\forall(\boldsymbol{\beta}, \alpha_i) \in \mathcal{B} \times \mathcal{A}$ by (A1). After having shown that that $\boldsymbol{\theta}_0$ is globally indentified, we only need to show that $\hat{Q}_T(\theta)$ converges *uniformly* to $Q_0(\boldsymbol{\theta}_0)$ to ensure consistency. By (A4) the objective function, $Q_0(\theta)$ is continuous on Θ. By (A1), $\mathbf{y}_t|\mathcal{F}_{t-1} \sim i.i.d$, and (A4), $\mathbb{E}\left[l_t(\boldsymbol{\theta})\right] < \infty$, so we can employ Kolmogorov's strong law of large numbers:

$$T^{-1}\sum_{t=1}^{T}\left(\sum_{i=1}^{N} log\, f(y_{i,t}|\mathcal{F}_{t-1};\boldsymbol{\beta},\alpha_i)\right) \xrightarrow{p} \mathbb{E}\left[\sum_{i=1}^{N} log\, f(y_{i,t}|\mathcal{F}_{t-1};\boldsymbol{\beta},\alpha_i)\right] \; as\; T\rightarrow\infty$$

Futhermore, as $\hat{Q}_T(\boldsymbol{\theta})$ is stochastically equicontinuous on Θ by (A5) and that Θ is compact by (A3) we know from Theorem 1 in Andrews (1992) that $Q_0(\boldsymbol{\theta}_0)$ converges *uniformly* to $Q_0(\boldsymbol{\theta}_0)$. □

Assumptions (continued):

(**A.6**) $\boldsymbol{\theta}_0 \in interior(\Theta)$.

(**A.7**) $l_t(\boldsymbol{\theta})$ is a twice differentiable function in $\boldsymbol{\theta}$.

(**A.8**) For some neighborhood $\mathcal{N}$ of $\boldsymbol{\theta}_0$:$\mathbb{E}\left[\sup_{\theta\in\mathcal{N}} \|\nabla_{\theta\theta} l_t(\boldsymbol{\theta})\|\right] < \infty$ such that $T^{-1}\sum_{t=1}^{T}\nabla_{\theta\theta} l_t(\boldsymbol{\theta}) \rightarrow E[\nabla_{\theta\theta} l_t(\boldsymbol{\theta}_0)]$.

(**A.9**) $\mathbb{E}[\nabla_{\theta\theta} l_t(\boldsymbol{\theta})]$ exist and is non-singular.

Proposition 2. *Let $\hat{\boldsymbol{\theta}}_{PL}$ be the partial likelihood estimator defined in* (2.3.1) *on page 60. If the assumptions of proposition 1 and (A6-A9) hold then:*

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_{PL}-\boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}(0,\mathbf{H}_0^{-1}\mathbf{J}_0\mathbf{H}_0^{-1})$$

where $\mathbf{J}_0$ *and* $\mathbf{H}_0$ *are defined as:*

$$\mathbf{J}_0 = \mathbb{E}\left[\begin{pmatrix} \sum_{i=1}^{N}\frac{[y_{i,t}-G(a_{i,t}(\boldsymbol{\theta}_i))]G'(a_{i,t}(\boldsymbol{\theta}_i))}{G(a_{i,t}(\boldsymbol{\theta}_i))(1-G(a_{i,t}(\boldsymbol{\theta}_i)))}\mathbf{x}_{i,t-1} \\ \frac{(y_{1,t}-G(a_{1,t}(\boldsymbol{\theta}_1)))G'(a_{1,t}(\boldsymbol{\theta}_1))}{G(a_{1,t}(\boldsymbol{\theta}_1))(1-G(a_{1,t}(\boldsymbol{\theta}_1)))} \\ \vdots \\ \frac{(y_{N,t}-G(a_{N,t}(\boldsymbol{\theta}_N)))G'(a_{N,t}(\boldsymbol{\theta}_N))}{G(a_{N,t}(\boldsymbol{\theta}_N))(1-G(a_{N,t}(\boldsymbol{\theta}_N)))} \end{pmatrix}\begin{pmatrix} \sum_{i=1}^{N}\frac{[y_{i,t}-G(a_{i,t}(\boldsymbol{\theta}_i))]G'(a_{i,t}(\boldsymbol{\theta}_i))}{G(a_{i,t}(\boldsymbol{\theta}_i))(1-G(a_{i,t}(\boldsymbol{\theta}_i)))}\mathbf{x}_{i,t-1} \\ \frac{(y_{1,t}-G(a_{1,t}(\boldsymbol{\theta}_1)))G'(a_{1,t}(\boldsymbol{\theta}_1))}{G(a_{1,t}(\boldsymbol{\theta}_1))(1-G(a_{1,t}(\boldsymbol{\theta}_1)))} \\ \vdots \\ \frac{(y_{N,t}-G(a_{N,t}(\boldsymbol{\theta}_N)))G'(a_{N,t}(\boldsymbol{\theta}_N))}{G(a_{N,t}(\boldsymbol{\theta}_N))(1-G(a_{N,t}(\boldsymbol{\theta}_N)))} \end{pmatrix}^T\right]$$

$$\mathbf{H}_0 = \mathbb{E}\begin{bmatrix} \sum_{i=1}^{N} q_{i,t}(\boldsymbol{\theta}_{i,0})\mathbf{x}_{i,t-h}\mathbf{x}_{i,t-1}^T & q_{1,t}(\boldsymbol{\theta}_{1,0})\mathbf{x}_{1,t-1} & \cdots & q_{N,t}(\boldsymbol{\theta}_{N,0})\mathbf{x}_{N,t-1} \\ q_{1,t}(\boldsymbol{\theta}_{1,0})\mathbf{x}_{1,t-1}^T & q_{1,t}(\boldsymbol{\theta}_{1,0}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ q_{N,t}(\boldsymbol{\theta}_{N,0})\mathbf{x}_{N,t-1}^T & 0 & \cdots & q_{N,t}(\boldsymbol{\theta}_{N,0}) \end{bmatrix}$$

Proof. In order to derive the asymptotic distribution of the estimator we require an asymptotically valid closed form representation for $\sqrt{T}\left(\hat{\boldsymbol{\theta}}_{PL}-\boldsymbol{\theta}_0\right)$. Generally, there does not exist a closed form solution for $\hat{\boldsymbol{\theta}}_{PL}$ but, as $l_t(\boldsymbol{\theta})$ is a twice differentiable function by (A7), we can employ a mean-value expansion around $\nabla_\theta \hat{Q}_T(\hat{\boldsymbol{\theta}}_{PL}) = T^{-1}\sum_{t=1}^{T}\nabla_\theta l_t(\hat{\boldsymbol{\theta}}_{PL})$:

$$\nabla_\theta \hat{Q}_T(\hat{\boldsymbol{\theta}}_{PL}) = \nabla_\theta \hat{Q}_T(\boldsymbol{\theta}_0) + \nabla_{\theta\theta}\hat{Q}_T(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{PL}-\boldsymbol{\theta}_0)$$

where $\tilde{\boldsymbol{\theta}} = diag(\boldsymbol{\lambda})\boldsymbol{\theta}_0 + (\boldsymbol{i} - diag(\boldsymbol{\lambda}))\hat{\boldsymbol{\theta}}_{PL}$, $\boldsymbol{\lambda} \in [0,1]^{k+N}$ and

$$\nabla_\theta l_t(\boldsymbol{\theta}) = \begin{pmatrix} \sum_{i=1}^{N} \frac{[y_{i,t}-G(a_{i,t}(\boldsymbol{\theta}_i))]G'(a_{i,t}(\boldsymbol{\theta}_i))}{G(a_{i,t}(\boldsymbol{\theta}_i))(1-G(a_{i,t}(\boldsymbol{\theta}_i)))}\mathbf{x}_{i,t-h} \\ \frac{(y_{1,t}-G(a_{1,t}(\boldsymbol{\theta}_1)))G'(a_{1,t}(\boldsymbol{\theta}_1))}{G(a_{1,t}(\boldsymbol{\theta}_1))(1-G(a_{1,t}(\boldsymbol{\theta}_1)))} \\ \vdots \\ \frac{(y_{N,t}-G(a_{N,t}(\boldsymbol{\theta}_N)))G'(a_{N,t}(\boldsymbol{\theta}_N))}{G(a_{N,t}(\boldsymbol{\theta}_N))(1-G(a_{N,t}(\boldsymbol{\theta}_N)))} \end{pmatrix}$$

From equation (2.3.1), and (A6), we know that the partial likelihood estimator satisfies $\nabla_\theta \hat{Q}_T(\hat{\boldsymbol{\theta}}_{PL}) = 0$. This leads to the following expression:

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_{PL}-\boldsymbol{\theta}_0) = -(T^{-1}\nabla_{\theta\theta}\hat{Q}_T(\tilde{\boldsymbol{\theta}}))^{-1}T^{-0.5}\nabla_\theta \hat{Q}_T(\hat{\boldsymbol{\theta}}_0).$$

We first analyse the asymptotic properties of $T^{-1}\nabla_{\theta\theta}\hat{Q}_T(\tilde{\boldsymbol{\theta}}) = T^{-1}\sum_{t=1}^{T}\nabla_{\theta\theta}l_t(\tilde{\boldsymbol{\theta}})$. Firstly, it is easy to verify that:

$$\nabla_{\theta\theta}l_t(\boldsymbol{\theta}) = \begin{bmatrix} \sum_{i=1}^{N}\tilde{q}_{i,t}(\boldsymbol{\theta}_i)\mathbf{x}_{i,t-1}\mathbf{x}_{i,t-h}^T & \tilde{q}_{1,t}(\boldsymbol{\theta}_1)\mathbf{x}_{1,t-1} & \cdots & \tilde{q}_{N,t}(\boldsymbol{\theta}_N)\mathbf{x}_{N,t-1} \\ \tilde{q}_{1,t}(\boldsymbol{\theta}_1)\mathbf{x}_{1,t-1}^T & \tilde{q}_{1,t}(\boldsymbol{\theta}_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{q}_{N,t}(\boldsymbol{\theta}_N)\mathbf{x}_{N,t-1}^T & 0 & \cdots & \tilde{q}_{N,t}(\boldsymbol{\theta}_N) \end{bmatrix}$$

By Proposition 1, $\hat{\boldsymbol{\theta}}_{PL} \to \boldsymbol{\theta}_0\, as\, T \to \infty$ which implies that $\tilde{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$ *as* $T \to \infty$. Furthermore, by (A1) and (A8) we have that $T^{-1}\sum_{t=1}^{T}\nabla_{\theta\theta}l_t(\tilde{\boldsymbol{\theta}}) \xrightarrow{p} E[\nabla_{\theta\theta}l_t(\boldsymbol{\theta}_0)]$. By the law of iterated expectations, the expectation of the first part of $\tilde{q}_{i,t}(\boldsymbol{\theta}_0)$ is zero:

$$\mathbb{E}\left[\frac{y_{i,t}-G(a_{i,t}(\boldsymbol{\theta}_i))}{G(a_{i,t}(\boldsymbol{\theta}_i))(1-G(a_{i,t}(\boldsymbol{\theta}_i)))}G''(a_{i,t}(\boldsymbol{\theta}_i))\right] = \mathbb{E}\left[\frac{G''(a_{i,t}(\boldsymbol{\theta}_i))}{G(a_{i,t}(\boldsymbol{\theta}_i)^0_{i,t})(1-G(a_{i,t}(\boldsymbol{\theta}_i)))}\left(\mathbb{E}\left[y_{i,t}\,|\mathcal{F}_{t-1}\right]-G(a^0_{i,t})\right)\right] = 0$$

as $a_{i,t}(\boldsymbol{\theta}_i) \in \mathcal{F}_{t-1}$. Therefore,

$$\begin{aligned}\mathbf{H}_0 &= \mathbb{E}[\nabla_{\theta\theta} l_t(\hat{\boldsymbol{\theta}}_0)] \\ &= -\mathbb{E}\begin{bmatrix} \sum_{i=1}^N q_{i,t}(\boldsymbol{\theta}_i)\mathbf{x}_{i,t-1}\mathbf{x}_{i,t-h}^T & q_{1,t}(\boldsymbol{\theta}_{1,0})\mathbf{x}_{1,t-1} & \cdots & q_{N,t}(\boldsymbol{\theta}_{N,0})\mathbf{x}_{N,t-1} \\ q_{1,t}(\boldsymbol{\theta}_{1,0})\mathbf{x}_{1,t-1}^T & q_{1,t}(\boldsymbol{\theta}_{1,0}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ q_{N,t}(\boldsymbol{\theta}_{N,0})\mathbf{x}_{N,t-1}^T & 0 & \cdots & q_{N,t}(\boldsymbol{\theta}_{N,0}) \end{bmatrix}\end{aligned}$$

Furthermore, as $\mathbf{H}_0$ is invertible by (A9), we get that $(T^{-1}\nabla_{\theta\theta}\hat{Q}_T(\tilde{\boldsymbol{\theta}}))^{-1} \xrightarrow{p} \mathbf{H}_0^{-1}$ by the continuous mapping theorem.

We now analyse the asymptotic properties of $T^{-0.5}\nabla_\theta \hat{Q}_T(\boldsymbol{\theta}_0)$. By (A1), we can apply the Lindeberg–Lévy Central Limit Theorem:

$$T^{-0.5}\nabla_\theta \hat{Q}_T(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{J}_0)$$

where

$$\begin{aligned}\mathbf{J}_0 &= E[\nabla_\theta l_t(\boldsymbol{\theta}_0)(\nabla_\theta l_t(\boldsymbol{\theta}_0))^T] \\ &= \mathbb{E}\left[\begin{pmatrix} \sum_{i=1}^N \frac{(y_{i,t}-G(a^0_{i,t}))G'(a^0_{i,t})}{G(a^0_{i,t})(1-G(a^0_{i,t}))}\mathbf{x}_{i,t-1} \\ \frac{(y_{1,t}-G(a^0_{1,t}))G'(a^0_{1,t})}{G(a^0_{1,t})(1-G(a^0_{1,t}))} \\ \vdots \\ \frac{(y_{N,t}-G(a^0_{N,t}))G'(a^0_{N,t})}{G(a^0_{N,t})(1-G(a^0_{N,t}))} \end{pmatrix}\begin{pmatrix} \sum_{i=1}^N \frac{(y_{i,t}-G(a^0_{i,t}))G'(a^0_{i,t})}{G(a^0_{i,t})(1-G(a^0_{i,t}))}\mathbf{x}_{i,t-1} \\ \frac{(y_{1,t}-G(a^0_{1,t}))G'(a^0_{1,t})}{G(a^0_{1,t})(1-G(a^0_{1,t}))} \\ \vdots \\ \frac{(y_{N,t}-G(a^0_{N,t}))G'(a^0_{N,t})}{G(a^0_{N,t})(1-G(a^0_{N,t}))} \end{pmatrix}^T\right]\end{aligned}$$

where we have utilized that by (A6), $\boldsymbol{\theta}_0 \in \overset{\circ}{\Theta}$, so $E[\nabla_\theta l_t(\boldsymbol{\theta}_0)] = 0$. Combining these two results leads to the desired result: $\sqrt{T}(\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{H}_0^{-1}\mathbf{J}_0\mathbf{H}_0^{-1})$. $\square$

Proof of Remark: If $Cov[y_{i,t}, y_{j,t}|\mathcal{F}_{t-1}] = 0\,\forall i \neq j$ then $avar(\hat{\boldsymbol{\theta}}_{PL}) = \mathbf{H}_0^{-1}$. This condition implies, by the law of iterated expectations, that the cross products in $\mathbf{J}_0$ cancels out:

$$E\left[\frac{[y_{i,t} - G(a_{i,t}(\boldsymbol{\theta}_{i,0}))]G'(a_{i,t}(\boldsymbol{\theta}_{i,0}))}{G(a_{i,t}(\boldsymbol{\theta}_{i,0}))(1 - G(a_{i,t}(\boldsymbol{\theta}_{i,0})))}\mathbf{x}_{i,t-1}\frac{[y_{j,t} - G(a_{j,t}(\boldsymbol{\theta}_{i,0}))]G'(a_{j,t}(\boldsymbol{\theta}_{i,0}))}{G(a_{j,t}(\boldsymbol{\theta}_{i,0}))(1 - G(a_{j,t}(\boldsymbol{\theta}_{i,0})))}\mathbf{x}_{j,t-1}\,|\mathcal{F}_{t-1}\right] =$$

$$E\left[\{y_{i,t} - G(a_{j,t}(\boldsymbol{\theta}_j))\}\{y_{j,t} - G(a_{j,t}(\boldsymbol{\theta}_j))\}\,|\mathcal{F}_{t-1}\right] \times \left[\frac{G'(a_{i,t}(\boldsymbol{\theta}_i))G'(a_{j,t}(\boldsymbol{\theta}_j))\mathbf{x}_{i,t-1}\mathbf{x}_{j,t-1}^T}{G(a_{i,t}(\boldsymbol{\theta}_i))(1 - G(a_{i,t}(\boldsymbol{\theta}_i)))G(a_{j,t}(\boldsymbol{\theta}_j))(1 - G(a_{j,t}(\boldsymbol{\theta}_i)))}\right] =$$

$$\underbrace{Cov[y_{i,t}, y_{j,t}|\mathcal{F}_{t-1}]}_{0} \times \left[\frac{G'(a_{i,t}(\boldsymbol{\theta}_i))G'(a_{j,t}(\boldsymbol{\theta}_j))\mathbf{x}_{i,t-1}\mathbf{x}_{j,t-1}^T}{G(a_{i,t}(\boldsymbol{\theta}_i))(1 - G(a_{i,t}(\boldsymbol{\theta}_i)))G(a_{j,t}(\boldsymbol{\theta}_j))(1 - G(a_{j,t}(\boldsymbol{\theta}_i)))}\right] = 0$$

where we have used that $a_{i,t}(\boldsymbol{\theta}_i) \in \mathcal{F}_{t-1}\forall i$. This implies that

$$\begin{aligned}
\mathbf{J}_0 &= \mathbb{E}\left[\begin{pmatrix} \sum_{i=1}^N \frac{(y_{i,t}-G(a_{i,t}^0))G'(a_{i,t}^0)}{G(a_{i,t}^0)(1-G(a_{i,t}^0))}\mathbf{x}_{i,t-1} \\ \frac{(y_{1,t}-G(a_{1,t}^0))G'(a_{1,t}^0)}{G(a_{1,t}^0)(1-G(a_{1,t}^0))} \\ \vdots \\ \frac{(y_{N,t}-G(a_{N,t}^0))G'(a_{N,t}^0)}{G(a_{N,t}^0)(1-G(a_{N,t}^0))} \end{pmatrix}\begin{pmatrix} \sum_{i=1}^N \frac{(y_{i,t}-G(a_{i,t}^0))G'(a_{i,t}^0)}{G(a_{i,t}^0)(1-G(a_{i,t}^0))}\mathbf{x}_{i,t-1} \\ \frac{(y_{1,t}-G(a_{1,t}^0))G'(a_{1,t}^0)}{G(a_{1,t}^0)(1-G(a_{1,t}^0))} \\ \vdots \\ \frac{(y_{N,t}-G(a_{N,t}^0))G'(a_{N,t}^0)}{G(a_{N,t}^0)(1-G(a_{N,t}^0))} \end{pmatrix}^T\right] \\
&= \mathbb{E}\begin{bmatrix} \sum_{i=1}^N q_{i,t}(\boldsymbol{\theta}_i)\mathbf{x}_{i,t-1}\mathbf{x}_{i,t-1}^T & q_{1,t}(\boldsymbol{\theta}_{1,0})\mathbf{x}_{1,t-1} & \cdots & q_{N,t}(\boldsymbol{\theta}_{N,0})\mathbf{x}_{N,t-1} \\ q_{1,t}(\boldsymbol{\theta}_{1,0})\mathbf{x}_{1,t-1}^T & q_{1,t}(\boldsymbol{\theta}_{1,0}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ q_{N,t}(\boldsymbol{\theta}_{N,0})\mathbf{x}_{N,t-1}^T & 0 & \cdots & q_{N,t}(\boldsymbol{\theta}_{N,0}) \end{bmatrix} \\
&= -\mathbf{H}_0
\end{aligned}$$

2.8 Bibliography

Allen, F. & Gale, D. (2000*a*), 'Bubbles and crises', *The Economic Journal* **110**(460), 236–255.

Allen, F. & Gale, D. (2000*b*), 'Financial contagion', *Journal of political economy* **108**(1), 1–33.

Andrews, D. (1992), 'Generic uniform convergence', *Econometric Theory* **8**(02), 241–257.

Bandt, O., Hartmann, P. & Peydró, J. (2009), *Systemic Risk in Banking: An Update*, Oxford Handbook of Banking, ed. by A. Berger, P. Molyneux and J. Wilson, Oxford University Press.

Bell, J. & Pain, D. (2000), 'Leading indicator models of banking crises–a critical review', *Financial Stability Review, Bank of England* **9**, 113–129.

Borio, C. & Drehmann, M. (2009), 'Assessing the risk of banking crises–revisited', *BIS Quarterly Review* **March**, 29–46.

Borio, C. & Lowe, P. (2002), 'Assessing the risk of banking crises', *BIS Quarterly Review* **December**, 43–54.

Davis, E. & Karim, D. (2008), 'Comparing early warning systems for banking crises', *Journal of Financial Stability* **4**(2), 89–120.

Demirgüç-Kunt, A. & Detragiache, E. (1998*a*), 'The determinants of banking crises in developing and developed countries', *International Monetary Fund Staff Papers* **45**(1), 81–109.

Demirgüç-Kunt, A. & Detragiache, E. (1998*b*), 'financial liberalization and financial fragility'. IMF Working Paper No. WP/98/83, International Monetary Fund.

Demirgüç-Kunt, A. & Detragiache, E. (2000), 'Monitoring banking sector fragility: a multivariate logit approach', *The World Bank Economic Review* **14**(2), 287–307.

Demirgüç-Kunt, A. & Detragiache, E. (2002), 'Does deposit insurance increase banking system stability? an empirical investigation', *Journal of Monetary Economics* **49**(7), 1373–1406.

Demirgüç-Kunt, A. & Detragiache, E. (2005), 'Cross-country empirical studies of systemic bank distress: A survey', *National Institute Economic Review* **192**(1), 68–83.

Diamond, D. & Dybvig, P. (1983), 'Bank runs, deposit insurance, and liquidity', *The Journal of Political Economy* **91**(3), 401–419.

Domac, I. & Martinez Peria, M. (2003), 'Banking crises and exchange rate regimes: is there a link?', *Journal of International Economics* **61**(1), 41–72.

Drehmann, M., Borio, C. & Tsatsaronis, K. (2011), 'Anchoring countercyclical capital buffers: The role of credit aggregates', *International Journal of Central Banking* **7**(4), 189–240.

Gaytán, A. & Johnson, C. (2002), 'a review of the literature on early warning systems for banking crises'. Working Paper no. 183, Central Bank of Chile.

Goldstein, M., Kaminsky, G. & Reinhart, C. (2000), 'assessing financial vulnerability: an early warning system for emerging markets'. Institute for International Economics, Washington D.C.

González-Hermosillo, B. (1996), 'banking sector fragility and systemic sources of fragility'. IMF Working Paper No. WP/96/12, International Monetary Fund.

González-Hermosillo, B. (1999), 'determinants of ex-ante banking system distress: A macro-micro empirical exploration of some recent episodes'. IMF Working Paper No. WP/99/33, International Monetary Fund.

Hahm, J., Shin, H. & Shin, K. (2011), 'non-core bank liabilities and financial vulnerability'. Working Paper, Princeton University.

Hausman, J. A. (1978), 'Specification tests in econometrics', *Econometrica: Journal of the Econometric Society* **46**(6), 1251–1271.

Heckman, J. (1981), 'the incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process and some monte carlo evidence'. in Manski, C. and D. McFadden, eds., Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge, pp. 114-178.

Hodrick, R. & Prescott, E. (1997), 'Postwar us business cycles: an empirical investigation', *Journal of Money, Credit, and Banking* **29**, 1–16.

Hutchison, M. & McDill, K. (1999), 'are all banking crises alike? the japanese experience in international comparison'. Working Paper no. 7253, National Bureau of Economic Research.

IMF (1998), 'financial crises: causes and indicators'. World Economic Outlook, May, Washington DC.

IMF (2011), 'toward operationalizing macroprudential policies: When to act?'. Global Financial Stability Report, September, Washington DC.

Kaminsky, G. & Reinhart, C. (1999), 'The twin crises: the causes of banking and balance-of-payments problems', *American economic review* **89**(3), 473–500.

Kaminsky, G. & Reinhart, C. (2000), 'On crises, contagion, and confusion', *Journal of International Economics* **51**(1), 145–168.

Laeven, L. & Valencia, F. (2008), 'systemic banking crises: a new database'. IMF Working Paper No. WP/08/224, International Monetary Fund.

Laeven, L. & Valencia, F. (2010), 'resolution of banking crises: The good, the bad, and the ugly'. IMF Working Paper, No. WP/10/146, International Monetary Fund.

Lancaster, T. (2000), 'The incidental parameter problem since 1948', *Journal of econometrics* **95**(2), 391–413.

Newey, W. & McFadden, D. (1994), 'Large sample estimation and hypothesis testing', *Handbook of econometrics* **4**, 2111–2245.

Neyman, J. & Scott, E. (1948), 'Consistent estimates based on partially consistent observations', *Econometrica* **16**, 1–32.

Reinhart, C. M. & Rogoff, K. S. (2011), 'From financial crash to debt crisis', *The American Economic Review* **101**(5), 1676–1706.

Rüschendorf, L. (2009), 'On the distributional transform, sklar's theorem, and the empirical copula process', *Journal of Statistical Planning and Inference* **139**(11), 3921–3927.

Wong, J., Wong, T. & Leung, P. (2010), 'Predicting banking distress in the emeap economies', *Journal of Financial Stability* **6**(3), 169–179.

Chapter 3

Equity Premium Predictability: Weak Predictors, Shrinkage and Economic Value

The empirical finance literature has identified several financial variables that appear to predict the equity premium, in-sample. These findings do not necessarily violate efficient markets as a time-varying equity premium could arise, for example, as a result of countercyclical variation in investors' risk aversion. However, Welch & Goyal (2008) find that none of these variables have any predictive power out-of-sample. In this paper we show that the equity premium is predictable out-of-sample once you impose certain shrinkage restrictions on the model parameters. The approach is motivated by the observation that many of the proposed financial variables can be characterised as 'weak predictors'. This suggests that a James-Stein type estimator will provide a substantial reduction in risk relative to an unrestricted estimator. The out-of-sample explanatory power is small, but we show that it is, in fact,economically meaningful to an investor with time-invariant risk aversion. More specifically, depending on the investor's risk aversion, and whether short sales and borrowing are allowed, an investor with CARA preferences is willing to pay around 150 basis points per year for access to the information in the superior forecasts. Using a shrinkage decomposition we also provide some empirical evidence which suggests that standard combination forecast techniques tends to 'overshrink' the model parameters leading to suboptimal model forecasts.

Keywords: Equity Premium; Forecast Evaluation; James-Stein Estimator
Journal of Economic Literature Codes: C58; C53; G11.

3.1 Introduction

The empirical finance literature has identified several financial variables that appear to predict the equity premium, in-sample. Examples include the dividend-price ratio or the default spread. Importantly, this observation does not necessarily require abandonment of the traditional efficient markets paradigm. Campbell & Cochrane (1999), for example, has shown that rational investors with habit formation preferences might become more risk averse to volatility in consumption and wealth when the economy is weak, and vice versa, thus generating a countercyclical equilibrium equity premium. However, the empirical evidence of *out-of-sample* equity premium predictability has been questioned. Welch & Goyal (2008) has found that all of the proposed financial variables fail to consistently outperform a simple constant mean model in terms of forecasting accuracy. This finding has important implications as it suggests that none of these models would have helped an investor with constant risk aversion to profitably time the market in real time. This paper reassesses the empirical evidence of out-of-sample equity premium predictability. Motivated by the observation that the proposed financial variables can be characterised as 'weak predictors', we propose to estimate the models by a James-Stein type estimator. We find that this approach leads to a substantial improvement in forecasting accuracy and we reject the hypothesis that it has the same expected as a constant mean model. The out-of-sample explanatory power is relatively small, but we show that it is economically meaningful for a investor with time-invariant risk aversion. Depending on the investor's risk aversion, and whether short sales and borrowing are allowed, an investor is willing to pay up to 150 basis points annually in order to obtain the information in the superior forecast. Therefore, our findings suggest that the conclusions in Welch & Goyal (2008) are premature and could be the result of weak predictors rather than spurious equity premium predictability. We also show how a combination forecast approach, based on single predictor models, can be represented as a multivariate, kitchen sink model. Importantly, we find that such an approach tends to overshrink the model parameters leading to suboptimal model forecasts.

The rest of the paper is structured as follows. Section 2 provides a brief overview of the related literature. Section 3 describes the modelling framework and presents two statistical tests of equal unconditional predictive ability. Section 4 presents the statistical evidence for equity premium predictability, and this is followed by section 5 where we try to estimate the economic value of equity premium predictability. In section 6, we analyse in more detail how a small degree of predictability can translate into substantial utility gains. Finally, section 7 concludes the chapter.

3.2 Related Literature

The purpose of this section is to provide a brief overview of the literature on equity premium predicability. For a recent survey article see Campbell (2008), which this section also draws from.

Equity premium predictability is of great interest to both academics and practitioners and the topic has a long history. An early example can be seen in Dow (1920), which explored the role of dividend-price ratios in predicting the equity premium. The dividend-price ratio is an example of a broader class of indicators known as valuation ratios. A valuation ratio is in a way a 'natural' predictor as a strong valuation might indicate an undervalued stock. However, the use of valuation ratios did not carry much weight in the academic literature until Ball (1978) found that lagged values of the dividend-price ratio are positively correlated with asset returns, in-sample. Similar results were found for other valuation ratios, including: the earnings-price ratio (Campbell & Shiller (1988*b*) and Campbell & Shiller (1988*a*)), the book-to-market ratio (Kothari & Shanken 1997) and the dividend-payout ratio (Lamont 1998).

Economic variables related to macroeconomic risk have also been found to be correlated with the equity premium, including: inflation (Nelson 1976), long-term treasury bills (Hodrick 1992), the default spread (Keim & Stambaugh 1986), the term spread (Campbell 1987) and volatility (Guo 2006). More recently (Boudoukh, Michaely, Richardson & Roberts 2007) proposed that net equity expansion can be used as a predictor. Other recent studies that find evidence of in-sample predictability include Ang & Bekaert (2007), Lettau & Van Nieuwerburgh (2008) and Pástor & Stambaugh (2009).

All these empirical findings suggest that the equity premium is not a constant number that can be estimated more precisely over time, but rather it is an unobserved state variable whose value must be inferred at each point in time. At the same time, research in asset pricing theory has made financial economists more comfortable with the idea that the equity premium can change over time, even in an efficient market with rational investors. Therefore, a time-varying equity premium does not necessarily require abandonment of the traditional paradigm of financial economics for a behavioral or inefficient-markets alternative. For example, Campbell & Cochrane (1999) showed that rational investors with habit formation preferences might become more averse to volatility in consumption and wealth when the economy is weak, thus driving up the equilibrium equity premium. Therefore, both from an empirical and a theoretical perspective, the consensus view at the beginning of the new century appeared to be that the equity premium is predictable.

However, equity premium predictability remains controversial for two reasons. First, several authors have expressed concern that the apparent in-sample predictability of the equity premium, might be spurious. Many of the predictor variables in the literature are highly persistent and, as Stambaugh (1999) pointed out, that persistence leads to biased coefficients in predictive regressions if innovations in the predictor variable are correlated with returns such as valuation ratios. Recent literature discusses alternative econometric methods for correcting the Stambaugh bias and conducting valid inference. See, for example Campbell & Yogo (2006).

Second, some authors have questioned whether in-sample statistical significance is the right way to measure equity premium predictability. Goyal & Welch (2003) and Butler, Grullon & Weston (2005) look at out-of-sample stock market predictability and find little evidence in favour of genuine out-of-sample equity premium predictability. More recently, Welch & Goyal (2008) has conducted a comprehensive study of out-of-sample equity premium predictability and argues that any of the proposed financial variables would "have helped an investor with access only to available information to profitably time the market."

In this paper, we reassess the empirical evidence of out-of-sample equity premium predictability. We show that the equity premium is predictable out-of-sample once you impose certain shrinkage restrictions on the model parameters. The approach is motivated by the observation that many of the proposed financial variables can be characterised as weak predictors which in turn suggests that a James-Stein type estimator will provide a substantial improvement relative to an unrestricted estimator. We employ two tests to evaluate whether any of the models lead to a significantly lower expected loss than a simple constant mean model. The first test, due to McCracken (2007), ignores parameter estimation error and so is a test of whether the two models have the same expected loss. The second test, due to Giacomini & White (2006), acknowledge that the end-user of a forecast is forced to consume not only the 'true' forecast value, but the estimation error as well; therefore, this is actually a test of whether the two methods have the same expected loss.

Other studies have also reassessed the empirical evidence of out-of-sample equity premium predicability. Campbell & Thompson (2008) argue that a real world investor would not mechanically forecast using a linear regression, but would impose some restrictions on the regression coefficients. More specifically, they assume that the investor will not base his portfolio decision on the forecast if the regression coefficient has a different sign than what is expected based on a theoretical model, or if the predicted equity premium is negative. If both of these conditions are not satisfied, then the forecaster uses the historical average. Campbell and Thompson find that these restrictions substantially improve the out-of-sample performance of the models. However, one drawback to their paper is that they do not test whether the out-of-sample outperformance is statistically significant.

Rapach, Strauss & Zhou (2010) consider a combination forecast approach to the equity premium problem and find this approach delivers significant out-of-sample gains relative to a constant mean model, and it does so consistently over time. As we show in this paper, a combination forecast can be interpreted as a kitchen sink model estimated via a particular shrinked estimator. A forecast combination approach is therefore more natural when the econometrician is not estimating the forecast models herself. More importantly, we find that combination forecasts tend to overshrink the model parameters and that a James-Stein type estimator leads to more accurate equity premium forecasts.

Other important studies that reassess the evidence of equity premium predictability include Ferreira & Santa-Clara (2011) and Dangl & Halling (2012). Ferreira and Santa-Clara propose to forecast the three components of stock market returns (the dividend–price ratio, earnings growth and price–earnings ratio growth) separately and find that this improves the forecast accuracy. Dangl and Halling evaluate predictive regressions that explicitly consider the time-variation of coefficients in a Bayesian framework. They find that models with constant coefficients are dominated by models with time-varying coefficients.

3.3 Modelling Approach

In this section we define the modelling framework and we present two statistical tests of unconditional predictive ability, due to McCracken (2007) and Giacomini & White (2006). In order to evaluate the relative accuracy of the two forecasting models we need to take a stand on the choice of loss function that maps forecast errors into a utility loss for the forecast user. Therefore, we also derive an 'economic' loss function based on a parsimonious economic model describing an investor's optimal portfolio decision.

Our setup is as follows: let $\mathbf{Z} \equiv \{\mathbf{Z}_t : \Omega \rightarrow \mathbb{R}^{\mathrm{k}+1}, \mathrm{k} \in \mathbb{N}^*, \mathrm{t} = 1, ..., \mathrm{T}+1\}$ be a stochastic process defined on a complete probability space $(\Omega, \mathcal{F}, Pr)$, where $\mathcal{F} = \{\mathcal{F}_t, t = 1, ..., T+1\}$ and $\mathcal{F}_t$ is the $\sigma - field$ $\mathcal{F}_t \equiv \sigma\{\mathbf{Z}_s, s \leq t\}$. Let $\mathbf{Z}_t = (Y_t, \mathbf{X}_t)$, where $Y_t = R_t - R_0 \in \mathbb{R}$ denotes the excess equity return over the period $t-1$ to t, and $\mathbf{X}_t \in \mathbb{R}^{\mathrm{k}}$ denotes a $k \times 1$ vector of time t measurable financial variables. We will assume that the excess equity return can be described as

$$Y_{t+1} = \underbrace{\mathbb{E}[Y_{t+1}|\mathcal{F}_t]}_{Equity\ Premium} + \epsilon_{t+1} \ , \ \epsilon_{t+1}|\mathcal{F}_t \sim i.i.d.\, F(0, \sigma^2), \tag{3.3.1}$$

where $\mathbb{E}[\cdot|\mathcal{F}_t]$ denotes the conditional expectation operator, ϵ_{t+h} is an unobserved error term and F is an unspecified distribution with a mean of zero and variance given by σ^2. As Y_{t+1} denotes the excess equity return it is then clear that $\mathbb{E}[Y_{t+1}|\mathcal{F}_t]$ represents the equity premium. While our model specification allows the equity premium to vary through time, we will assume that the data is drawn from a stationary distribution such that $\mathbb{E}[Y_{t+1}|\mathcal{F}_t]$ is a time invariant function of $\mathcal{F}_t$. For ease of notation, we will use $\mu_{t+1,t} \equiv \mathbb{E}[Y_{t+1}|\mathcal{F}_t]$ to denote the equity premium going forward.

As the equity premium is unobserved, we will use the realised excess return, Y_{t+1}, as a proxy when evaluating the accuracy of a forecast. The excess return will be a conditionally unbiased estimate of the equity premium, by definition, but it will also contain a substantial amount of noise. Importantly, as highlighted in Hansen & Lunde (2006) and Patton (2011), the use of a conditionally unbiased, but imperfect, proxy can lead to undesirable outcomes in standard methods for comparing forecasts based on certain loss functions. However, the particular choice of loss function we will employ in this analysis, to be described in more detail below, will be 'robust' according to Patton's definition. This implies that the ranking of two forecasts, $\hat{\mu}^1_{t+1|t}$ and $\hat{\mu}^2_{t+1|t}$, by expected loss will be the same whether the ranking is done using the true conditional mean, $\mu_{t+1|t}$, or the conditionally unbiased proxy, Y_{t+1}. That is, $\mathbb{E}[\mathcal{L}(Y_{t+1}, \hat{\mu}^1_{t+1|t})] > \mathbb{E}[\mathcal{L}(Y_{t+1}, \hat{\mu}^2_{t+1|t})]$ implies that $\mathbb{E}[\mathcal{L}(\mu_{t+1|t}, \hat{\mu}^1_{t+1|t})] > \mathbb{E}[\mathcal{L}(\mu_{t+1|t}, \hat{\mu}^2_{t+1|t})]$ for a 'robust' loss function such as the one we employ.

The true model specification of the equity premium is unknown in practise and in general it is non-linear. In order to obtain an estimate of the unobserved equity premium we will employ a semi-parametric approach and assume that it can be represented by a parametric function, $\hat{\mu}_{t+1|t}(\boldsymbol{\beta}) = g(\mathcal{F}_t; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a finite dimensional parameter vector which is assumed to lie in the interior of the parameter space, Θ. When selecting a particular forecasting model one faces a tradeoff between approximation error and estimation error. In order to illustrate this it is useful to consider a decomposition of the forecast error based on the equity premium proxy, Y_{t+1}:

$$\begin{aligned}\hat{\epsilon}_{t+1} &= Y_{t+1} - \hat{\mu}_{t+1|t}(\hat{\boldsymbol{\beta}}) \\ &= \underbrace{(Y_{t+1} - \mu_{t+1|t})}_{Proxy\,Error} + \underbrace{(\mu_{t+1|t} - \hat{\mu}_{t+1|t}(\boldsymbol{\beta}_0))}_{Approximation\,Error} + \underbrace{(\hat{\mu}_{t+1|t}(\boldsymbol{\beta}_0) - \hat{\mu}_{t+1|t}(\hat{\boldsymbol{\beta}}))}_{Estimation\,Error}.\end{aligned}$$

The first term is the deviation of the proxy, Y_{t+1}, from the true equity premium. All forecasts, no matter how good, will suffer from this error because we use the realised excess return as a proxy for the equity premium. Consequently, forecast errors from different forecasts will be correlated by construction. The second term describes the error contribution caused by using an imperfect model of the conditional mean. The third and final term arises because the forecaster is forced to rely on an estimate of $\boldsymbol{\beta}_0$. The last two components of the forecast error generally entail a tradeoff when choosing a forecasting model. A simple parsimonious model will have a small estimation error, but it is likely that a larger model will provide a better approximation of the true conditional mean.

In line with previous research (e.g. Welch & Goyal (2008), Campbell & Thompson (2008) and Rapach et al. (2010)), we will assume that the equity premium can be represented by a linear model, $g(\mathcal{F}_t, \boldsymbol{\beta}) = \mathbb{E}[Y_{t+1}] + \boldsymbol{\beta}'(\mathbf{X}_t - \mathbb{E}[\mathbf{X}_t])$. As illustrated above, this assumption could potentially introduce approximation error into the forecast error. The no-predictability hypothesis translates into $\boldsymbol{\beta} = \mathbf{0}$. Consequently, we will use the forecasts from a constant mean model as a benchmark:

$$\hat{\mu}_{t+1|t}^{Benchmark} = \hat{\mathbb{E}}[Y_{t+1}].$$

Note that even if the equity premium is predictable, i.e. the constant mean model is misspecified, the constant mean model may still do a reasonable job. The reason is that while the benchmark forecast may have a relatively large amount of approximation error due to its simplicity, it will have a minimal amount of estimation error due to the limited number of parameters. In fact, by construction, it will have a lower estimation error than any of the other models that we will consider. Therefore, the success of a particular prediction model depends on whether the increase in estimation error will be outweighed by a potential reduction in the approximation error.

All the different models that we will consider in this paper can be nested within a multivariate, kitchen sink model with shrinkage:

$$\hat{\mu}_{t+1|t}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \hat{\mathbb{E}}[Y_{t+1}] + \sum_{i=1}^{k}(1-\lambda_i)\beta_{i,t}(X_{i,t} - \hat{\mathbb{E}}[X_{i,t}]), \quad (3.3.2)$$

where λ_i denotes the degree of shrinkage for variable i. Clearly, this model nests the benchmark model for $\boldsymbol{\lambda} = \boldsymbol{\iota}$. Welch & Goyal (2008) construct forecasts from an unrestricted kitchen sink model, $\hat{\mu}^{KS}_{t+1|t}(\hat{\boldsymbol{\beta}}_{OLS,t}, \boldsymbol{\lambda} = \mathbf{0})$, where

$$\hat{\boldsymbol{\beta}}_{OLS,t} = (\Sigma^{t}_{j=t-R+1}\tilde{\mathbf{X}}_{t-1}\tilde{\mathbf{X}}'_{t-1})^{-1}(\Sigma^{t}_{j=t-R+1}\tilde{\mathbf{X}}_{t-1}\tilde{Y}_t), \quad (3.3.3)$$

$\tilde{\mathbf{X}}_t = (\mathbf{X}_t - \bar{\mathbf{X}})$ and $\tilde{Y}_t = Y_t - \bar{Y}$, and find that it performs very poorly in terms of out-of-sample forecast accuracy. At first glance this is a surprising result as this very flexible model might provide an accurate approximation of the true equity premium. However, while this intuition is correct, it is also likely that the estimation error, due to the high number of parameters, will be very large and this could be driving the poor performance of the model. Welch & Goyal (2008) also consider k single predictor forecasting models. These models are also nested within the model above as $\hat{\mu}^{i}_{t+1|t}(\mathbf{e}_i\hat{\boldsymbol{\beta}}_{SPE,t}, \boldsymbol{\lambda} = \mathbf{0})$ where

$$\hat{\boldsymbol{\beta}}_{SPE,t} = (diag(\Sigma^{t}_{j=t-R+1}\tilde{\mathbf{X}}_{t-1}\tilde{\mathbf{X}}'_{t-1}))^{-1}(\Sigma^{t}_{j=t-R+1}\tilde{\mathbf{X}}_{t-1}\tilde{Y}_t), \quad (3.3.4)$$

and $\mathbf{e}_i$ is a $k \times k$ diagonal matrix with 1 in the $i'th$ diagonal and zero otherwise. The SPE subscript is an abbreviation for single-predictor-estimator as the estimator restricts the explanatory variables to be uncorrelated such that it equals the individual estimators from a single predictor model. Interestingly, Goyal and Welch find that none of the k single predictor models significantly outperforms the constant mean model in terms out-of-sample forecast accuracy. Therefore, their findings question whether the equity premium is predicable out-of-sample and argue that “these models would not have helped an investor with access only to available information to profitably time the market”.

Rapach, Strauss & Zhou (2010) propose to use a combination forecast approach based on single predictor model forecasts, $\sum_{i=1}^{k}\omega_{i,t}\hat{\mu}^{i}_{t+1|t}(\mathbf{e}_i\hat{\boldsymbol{\beta}}_{SPE,t}, \mathbf{0})$. If the weights sum to unity, i.e.

if $\boldsymbol{\omega}'\boldsymbol{\iota} = 1$, we can describe the combination forecast as $\hat{\mu}^{CF}_{t+1|t}(\hat{\boldsymbol{\beta}}_{SPE,t}, \boldsymbol{\iota} - \boldsymbol{\omega}_t)$. Therefore, the combination forecast is simply equal to the unrestricted kitchen sink model forecast,

$$\hat{\mu}^{CF}_{t+1|t}(\hat{\boldsymbol{\beta}}_{SPE,t}, \boldsymbol{\lambda} = \boldsymbol{\iota} - \boldsymbol{\omega}) = \hat{\mu}^{KS}_{t+1|t}(\hat{\boldsymbol{\beta}}_{OLS,t}, \mathbf{0}) - \underbrace{\sum_{i=1}^{k}(1 - \omega_{i,t})\beta^{OLS}_{i,t}\tilde{X}_{i,t}}_{Shrinkage} + \underbrace{\sum_{i=1}^{k}\omega_i(\hat{\beta}^{SPE}_{i,t} - \hat{\beta}^{OLS}_{i,t})\tilde{X}_{i,t}}_{Bias\,Correction}, \tag{3.3.5}$$

with shrinkage and bias-correction terms added.[1] The second component is denoted bias correction as the $\hat{\boldsymbol{\beta}}_{SPE,t}$ estimator generally is biased due to the omitted variables.

Now, we will consider different combination forecast approaches. Firstly, we will use the static forecast combination methods: mean, trimmed mean and median. The mean combination set $\omega_{i,t} = \frac{1}{k}$ for all i, t, the trimmed mean set $\omega_{i,t} = 0$ for the forecast with the highest and smallest value, and $\omega_{i,t} = \frac{1}{k-2}$ for the remaining forecasts, for all t. The median combination forecast is the median of $\hat{\mu}^{1}_{t+1|t}(\mathbf{e}_1\hat{\boldsymbol{\beta}}_{SPE,t}, \boldsymbol{\lambda} = \mathbf{0})$,...,$\hat{\mu}^{k}_{t+1|t}(\mathbf{e}_k\hat{\boldsymbol{\beta}}_{SPE,t}, \boldsymbol{\lambda} = \mathbf{0})$, for all t. Other combination forecast methods attempts to estimate the optimal combination forecast based on realised forecasts in a holdout period. For example, one approach is to choose the combination that minimises the squared forecast error. If the out-of-sample forecast errors, $\hat{\boldsymbol{e}}_{t+1} = (\hat{e}^1_{t+1}, ..., \hat{e}^k_{t+1})'$ where $\hat{e}^i_{t+1} = Y_{t+1} - \hat{\mu}^i_{t+1|t}(\mathbf{e}_i\hat{\boldsymbol{\beta}}_{SPE,t}, \mathbf{0})$, have finite second moments we can compute the optimal weights as the solution to $\underset{\boldsymbol{\omega}_t \in \mathbb{R}^{\mathrm{N}}}{arg\,min}\,\mathbb{E}\left[(\boldsymbol{\omega}_t'\hat{\boldsymbol{\epsilon}}_{t+1})^2|\mathcal{F}_t\right]$ where we again have imposed the condition that $\boldsymbol{\omega}_t'\iota = 1$. If all the forecasts are conditionally unbiased we can rewrite the objective function as $\boldsymbol{\omega}_t'\Omega\boldsymbol{\omega}_t$ where $\Omega = Var[\hat{\boldsymbol{\epsilon}}_{t+1}|\mathcal{F}_t]$. We know this problem from portfolio theory. It is identical to the problem of finding the global minimum-variance portfolio and the solution is given by $\boldsymbol{\omega}^*_t = \frac{\Omega^{-1}\iota}{\iota'\Omega^{-1}\iota}$. The optimal linear combination forecast will have an expected loss that is lower than the best individual forecast. However, Ω is unobservable and will have to be estimated by the data. This is often difficult in practise due to the accumulation of estimation errors when estimating a vast covariance matrix. Therefore, we will also consider an alternative

[1] Stock & Watson (2012) provides a simple shrinkage representation that describes the operational characteristics of various forecasting methods designed for a large number of orthogonal predictors (such as principal components). These methods include pretest methods, Bayesian model averaging, empirical Bayes, and bagging.

approach where we ignore the forecast errors' covariances. If we impose the condition that all the forecast errors are uncorrelated, $\tilde{\Omega} = diag(diag(\Omega))$, we get the following optimal combination $\omega^{**}_{i,t} = \frac{V[\hat{\epsilon}^i_{t+1}|\mathcal{F}^Z_t]}{\sum_{i=1}^N V[\hat{\epsilon}^i_{t+1}|\mathcal{F}^Z_t]^{-1}}$ $\forall i \in \{1, ..., N\}$. We will use the sample covariance matrix, based on a rolling window of 20 years, to estimate Ω.

The idea of combination forecasts goes back to Bates & Granger (1969) who showed that combinations of individual forecasts can outperform the individual forecasts themselves. For more details on the benefits of forecast combination techniques, see appendix D. The background is that decision makers often face multiple competing forecasts of the same economic variable. This could, for example, reflect differences in modelling approaches due to the fact that the true underlying data generating process is unknown. However, the combination forecast approach is less natural when the econometrician is estimating all the individual forecasting models herself. As we illustrated above, see 3.3.5, the econometrician could simply use a shrinked kitchen sink model estimated via a particular estimator.

In this paper, we argue that the equity premium is predictable but that the predictors are weak. That is, the magnitude of the parameters are small relative to the estimation error. This implies that the gains from smaller approximation error might be outweighted by a larger increase in the estimation error. Therefore, we propose to use a James-Stein estimator to estimate the model parameters in the multivariate, kitchen sink model in 3.3.2.

The ordinary least squares (OLS) estimator has many desirable properties in a conventional semi-parametric setting for a linear model. For example, under strict exogeneity and homoskedasticity, it is well known that the OLS estimator has the smallest variance among all linear unbiased estimators. However, if we give up unbiasedness, the OLS estimator does not have the smallest mean squared error. For example, shrinking the parameter estimates towards a particular restriction (e.g. zero) might result in a lower mean squared error. Clearly, this new shrinkage estimator will not be unbiased anymore, but what it pays for in bias it might make up for in variance. Stein (1956) first observed that an unconstrained Gaussian estimator is inadmissible when the dimension exceeds two; and shrinkage was introduced by James & Stein (1961) in the context of exact normal sampling. More specifically, suppose that we have an estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0 \in \Theta \subset \mathbb{R}^k$ which has the exact distribution $\hat{\boldsymbol{\beta}} \backsim \mathcal{N}(\boldsymbol{\beta}_0, \mathbf{V})$. The James-Stein estimator for $\boldsymbol{\beta}_0$ is

$$\hat{\boldsymbol{\beta}}_{JS} = \left(1 - \left(\frac{k-2}{\hat{\boldsymbol{\beta}}'\mathbf{V}^{-1}\hat{\boldsymbol{\beta}}}\right)\right)\hat{\boldsymbol{\beta}}.$$

When $k \geqslant 3$ the James-Stein estimator $\hat{\boldsymbol{\beta}}_{JS}$ has smaller risk than $\hat{\boldsymbol{\beta}}$ for a class of loss functions which includes weighted squared errors. An in-depth treatment of shrinkage theory can be found in Chapter 5 of Lehmann & Casella (1998). Importantly, the result above is a finite-sample so it is effectively restricted for econometric applications to the classic Gaussian (parametric) regression model with exogenous regressors. Hansen (2013) has recently extended James and Stein's methods to a broad array of conventional parametric econometric models. However, to the best of my knowledge, the results have still not been extended to a general, semi-parametric setting.

In our application, we will employ the James-Stein estimator to estimate the kitchen sink model,

$$\hat{\mu}^{JS}_{t+1|t}(\hat{\boldsymbol{\beta}}_{OLS,t}, \lambda_{JS}\boldsymbol{\iota}) = \mathbb{E}[Y_{t+1}] + (1-\lambda_{JS})\sum_{i=1}^{k}\hat{\beta}_{OLS,i}(X_{i,t} - \mathbb{E}[X_{i,t}]),$$

where $\lambda_{JS} = (k-2)(\hat{\boldsymbol{\beta}}'_{OLS,t}\hat{V}[\hat{\boldsymbol{\beta}}_{t,OLS}]^{-1}\hat{\boldsymbol{\beta}}_{OLS,t})^{-1}$. Importantly, Hansen shows formally that the risk reduction gained from using the James-Stein estimator is largest when the restriction is close to the true value. Here, where the restriction is zero, it simply means that the gains from the James-Stein type estimator are largest when the predictors are 'weak', meaning that the true coefficients are close to zero. However, as we have not made any distributional assumptions regarding the error term in our model specification in 3.3.1, i.e. since we employ a semi-parametric approach, we cannot expect to achieve risk reduction with certainty in our application. Furthermore, the feasible James-Stein estimator we will employ will be based on a estimate of the estimator's variance rather than the true variance.

3.3.1 Estimation Scheme

In order to evaluate a forecasting model's performance, out-of-sample, we will have to split the available sample into two pieces: a regression set of R observations and a subsequent prediction set of P observations. For each of the P observations in the out-of-sample set we can employ the forecast procedure as if we were actually in the position of forecasting out of sample, hereby constructing a series of forecasts $\{\hat{\mu}_{t+1|t}\}_{t=R}^{R+P-1}$. This is sometimes referred to as pseudo, real-time forecasting. Since each of the P forecasts are made at different points in time the forecaster has different amounts of information available at each forecast. This implies that, once we have decided which estimator to use, we also need to take a stand on how to estimate $\boldsymbol{\beta}$ over time. There are three commonly used estimation schemes: fixed, rolling and recursive. Under the fixed scheme the parameters are estimated once and the same point estimate is used for the entire out-of-sample period. In the rolling and recursive schemes the parameters are re-estimated every time a forecast is made. The recursive scheme utilises all past observations for the estimation, whereas the rolling scheme only uses a fixed window of the most recent observations. The recursive method is superior in the absence of structural breaks; however, in the presence of structural breaks the rolling estimation scheme adapts quicker to a new regime. In this study, we will use a rolling estimation scheme, as indicated by 3.3.3 and 3.3.4.

3.3.2 Forecast Evaluation and the Loss Function

In this chapter we are interested in evaluating the performance of a given forecasting model relative to a constant mean model. More specifically, we will statistically test the hypothesis that two models (or methods) have the same expected loss. Therefore, we need to take a stand on the choice of loss function. Here, rather than simply assuming an arbitrary loss function, we derive an economic loss function based on a parsimonious model describing a representative investor's optimal portfolio choice.

We assume that we are in a two-period world with only two assets: a riskless bond that yields a safe gross return of R_0 and a risky asset with a stochastic excess return of R_{t+1}.[2]

[2]The investor cares ultimately about consumption, not wealth. In this framework we implicitly assume that there is no consumption at time t, $C_t = 0$, such that savings equals wealth in the initial period. Since we are in a two period world it follows that $W_{t+1} = C_{t+1}$. Alternatively, we could also model the investor's saving decision. However, since we are only interested in the portfolio allocation, and not intertemporal consumption, this would make the model unnecessarily complicated.

We will make the following assumptions:

Assumption 1 (Utility Maximization) *The investor chooses his optimal portfolio, α_t^*, so as to maximize his expected utility in the second period,*

$$\alpha_t^* = \underset{\alpha_t \in \mathbb{R}}{arg\,max}\, \mathbb{E}[u(w_{t+1}(\alpha_t);\gamma)|\mathcal{F}_t]\,,$$

where $u(\cdot)$ denotes the investor's utility function, $w_{t+1}(\alpha_t) = \alpha_t R_{t+1} + (1-\alpha_t)R_0$ denotes the investor's wealth in the second period, γ is a parameter that determines the investor's risk aversion and $\mathcal{F}_t$ is a $\sigma - algebra$ that represents the investor's information set in the initial period.

Assumption 2 (Distribution of Risky Asset) *The risky return has the following conditional distribution:*

$$R_{t+1}|\mathcal{F}_t \sim \mathcal{N}(R_0 + \mu_{t+1|t}, \sigma^2),$$

where $\mu_{t+1|t}$ denotes the conditional equity premium, and σ^2 denotes the variance.

Assumption 3 (Investor Preferences) *The investor's preferences can be represented by a constant absolute risk aversion (CARA) utility function*

$$u(w_{t+1};\gamma) = -exp(-\gamma w_{t+1}),$$

where γ describes the degree of risk aversion.

Under these assumptions, in which we are describing the investor's optimal investment decision, we are now in a position where we can state a proposition describing the representative investor's loss function.

Proposition 1 (Loss Function): *Suppose that assumptions 1-3 above hold. Then, the investor's loss associated with forecasting the conditional mean with error can be described by the following loss function:*

$$\mathcal{L}(\mu_{t+1|t}, \hat{\mu}_{t+1|t}; \gamma, \sigma^2) \quad = \quad (2\gamma\sigma^2)^{-1}(\mu_{t+1|t} - \hat{\mu}_{t+1|t})^2.$$

Similarly, the representative investor's loss associated with forecasting the variance with error can be described by the following loss function:

$$\mathcal{L}(\sigma^2, \hat{\sigma}^2; \gamma, \mu_{t+1|t}) = \mu_{t+1|t}^2 \gamma^{-1} \left\{ \frac{1}{2\sigma^2} + \frac{\sigma^2}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^2} \right\}.$$

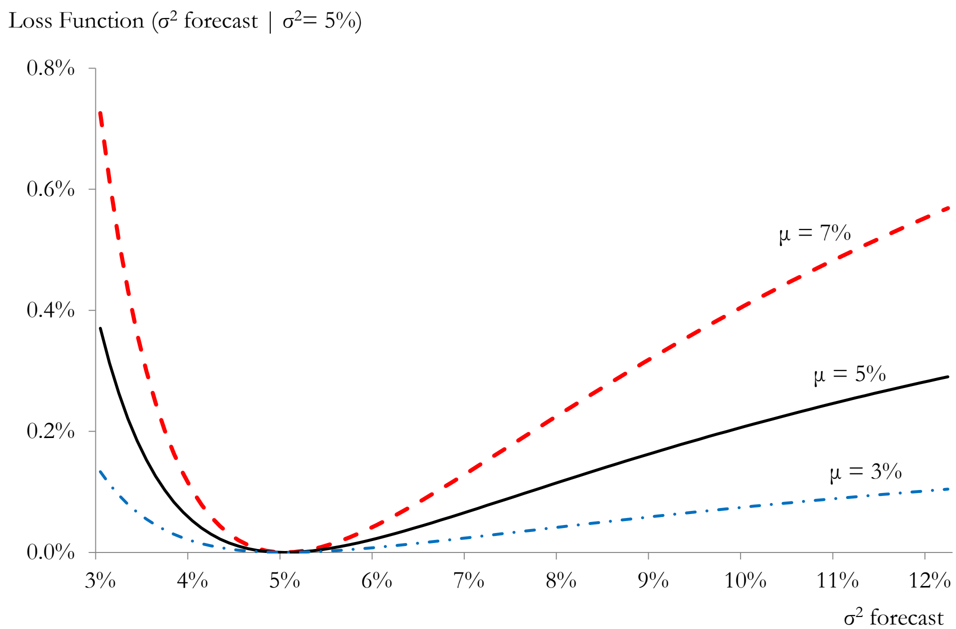
Proof: See section Appendix A. □

The loss functions describe the loss associated with forecasting the equity premium and volatility with error. Figure 3.1 illustrates the shape of the two loss functions. Interestingly, while there is an asymmetry in the loss function associated with volatility forecast errors, the loss function associated with forecasting the forecast errors in the equity premium has a mean squared error shape that is scaled by the inverse of the risky asset's variance.[3] If volatility is low then the loss is higher, all else equal, for a given equity premium. This reflects the fact that a larger fraction of the investor's wealth will be allocated to the risky asset and, therefore, that a forecast error in the equity premium will be more costly. Here, as the focus will be on evaluating equity premium forecasts, we will work under the assumption that the investor's loss can be approximated by squared error, $\mathcal{L}(\mu_{t+1|t}, \hat{\mu}_{t+1|t}) = (\mu_{t+1|t} - \hat{\mu}_{t+1|t})^2$. Therefore, even though it is widely documented that the conditional volatility of equity markets varies through time, we implicitly assume that the volatility is constant when evaluating the relative forecast accuracy.

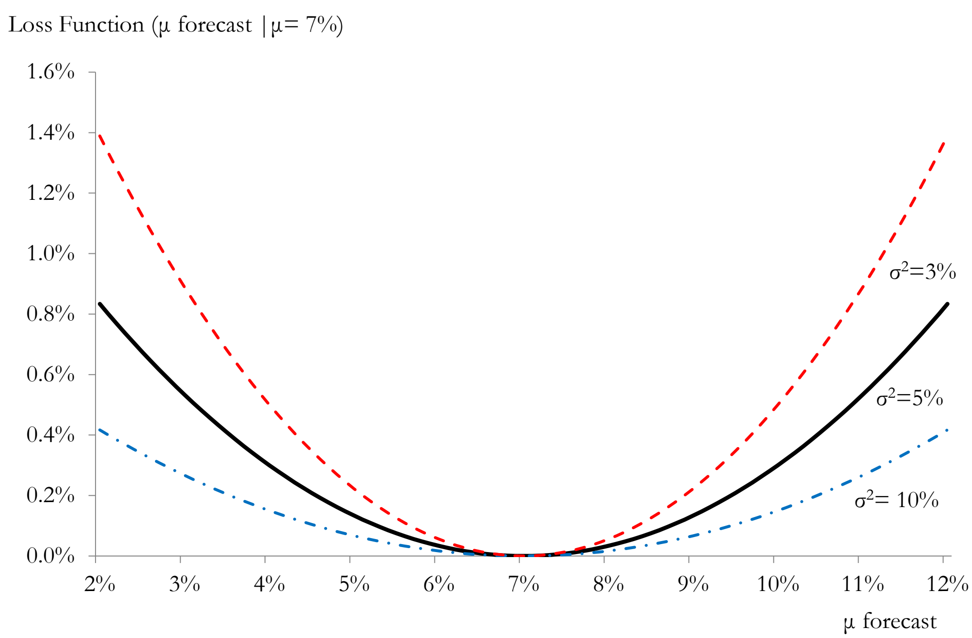
3.3.3 Tests of Unconditional Equal Predictive Ability

In this paper we are interested in evaluating the performance of a given forecasting model relative to the historical average. Specifically, we will test if two models have the same expected loss, due to West (1996) and McCracken (2007), and test if two methods have the same expected loss, due to Giacomini & White (2006). Both tests are based on the following

[3] West, Edison & Cho (1993) also find, in a slightly different framework, that a representative investor's loss is asymmetric in the volatility forecast.



(a) *Volatility loss function (for $\gamma = 3$)*



(b) *Conditional mean loss function (for $\gamma = 3$)*

Figure 3.1: *Loss function for the representative investor based on proposition 1*

test statistic:

$$OOS - t = \sqrt{P}\frac{\frac{1}{P}\sum_{t=R}^{R+P-1}(y_{t+1} - \hat{\mu}_{t+1|t}^{Benchmark})^2 - (y_{t+1} - \hat{\mu}_{t+1|t}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}))^2}{\sqrt{\hat{\sigma}_{LD}^2}},$$

where $\hat{\sigma}_{LD}^2$ denotes a consistent estimate of the asymptotic variance of the scaled loss difference. We will use a Newey-West type estimator to estimate this quantity.

Testing if Two Forecasting Models have the Same Expected Loss Suppose we want to test whether a particular model has the same expected loss as the historical average, $\mathcal{H}_0 : \mathbb{E}[(y_{t+1} - \mathbb{E}[y_t])^2] = \mathbb{E}[(y_{t+1} - \hat{\mu}_{t+1|t}(\boldsymbol{\beta}_0, \boldsymbol{\lambda}_0))^2]$, against the single sided alternative $\mathcal{H}_A : \mathbb{E}[(y_{t+1} - \mathbb{E}[y_t])^2] > \mathbb{E}[(y_{t+1} - \hat{\mu}_{t+1|t}(\boldsymbol{\beta}_0, \boldsymbol{\lambda}_0))^2]$. Note that the forecasts are evaluated at $\boldsymbol{\beta}_0$ (and $\boldsymbol{\lambda}_0$). When testing for equal predictive ability it is important to clarify whether the two models are nested or not. West (1996) shows that, under some regularity conditions, the $OOS - t$ test statistic is asymptotically normal, under the null, when the two models are not nested. However, all the models we consider nest the constant mean model so we cannot apply West's results. When the models are nested it is easy to show that the asymptotic variance in West (1996) is zero, but it is not immediate clear whether the test statistic is degenerate or divergent. In particular, it is not clear whether it is asymptotically normal. McCracken (2007) shows that under some conditions the test statistic follows a non-standard distribution when the two models are nested. Generally, the non-standard null distribution is shifted to the left relative to a standard normal distribution which reflects the fact that the estimation error has a higher impact on the larger model. The distribution depends on the limiting behaviour of the number of forecasts relative to the number of observations in the estimation period, $lim_{P,R\rightarrow\infty}\frac{P}{R} = \pi$, the number of additional variables in the largest model, k, and the estimation scheme. Appendix E contains a table with critical values for the asymptotic null distribution for different values of π and k based on a rolling estimation scheme. The critical values can be found by numerical methods.

An important limitation of this framework is that it does not work for non-parametric estimators. It also requires that the loss function used to evaluate the predictive ability is the same as that used to estimate the model parameters. In this context, where we assume a squared loss function, this implies that the parameters must be estimated using ordinary least squares. For further details see McCracken (2007).

Testing if Two Forecasting Methods Have the Same Expected Loss In the previous subsection we explained how to test whether two models have the same expected loss. The asymptotic distribution took into account that we are using an estimate of $\boldsymbol{\beta_0}$, rather than the true value, when we implement the test. However, a practical application of a forecast requires the end-user to consume not only the true forecast value, but the estimation error as well. Therefore, it is also interesting to test whether the two forecasting methods have the same expected loss. Giacomini & White (2006) show that under the null, $\mathcal{H}_0 : \mathbb{E}[(y_{t+1} - \overline{y_t})^2] = \mathbb{E}[(y_{t+1} - \hat{\mu}_{t+1|t}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}))^2]$, the $OOS - test$ statistic is asymptotically standard normal. However, it should be noted that the Giacomini and White framework only works under a rolling estimation scheme. To understand this you need to look at how the asymptotic distribution is derived. The asymptotics in West (1996) and McCracken (2007) are derived under conditions where $R \to \infty$, $P \to \infty$ and $P/R \to \pi \in [0, \infty)$. Thus the parameter estimation error is vanishing at a rate of $\sqrt{P/R}$. Giacomini & White (2006) change this framework such that $R \leq \bar{R}$ for some finite $\bar{R}$ and assumes that the forecaster is using a rolling estimation scheme. This implies that the estimation error is present in the limit.

3.4 Empirical Results

In this section we present the statistical evidence for out-of-sample equity premium predictability at a one month forecast horizon. We use monthly observations for 1926:12-2010:12 for the excess return on the value weighted CRSP. The realised monthly excess returns are illustrated in figure 3.2. The realised equity premium has been around 5-6% on average (annualised).

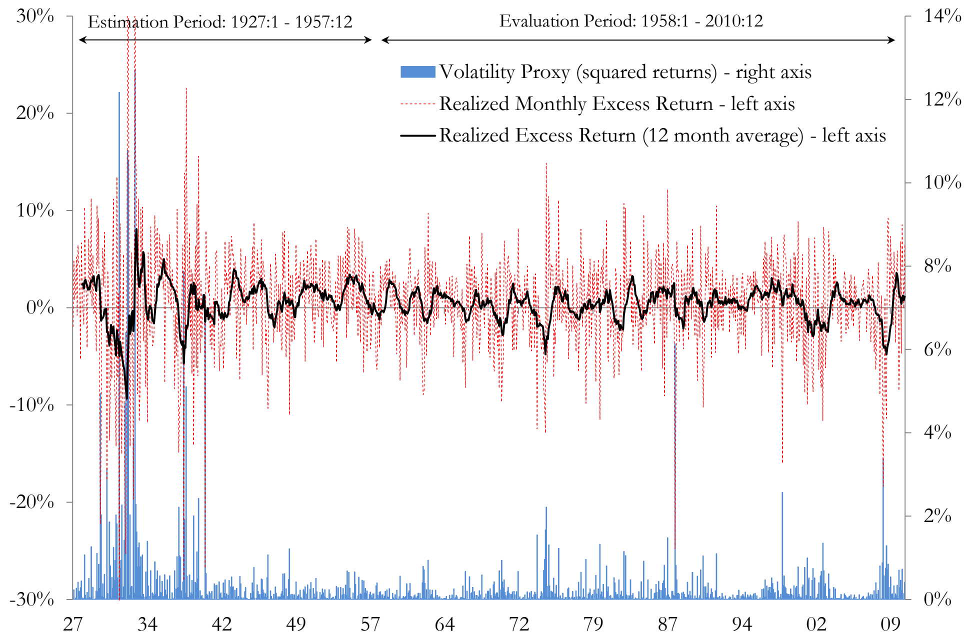


Figure 3.2: *Illustration of the monthly excess return on the CRSP index: 1927:1-2010:12.*

We will construct different equity premium forecasts based on the following eleven financial variables: dividend-price ratio, earnings-price ratio, book-to-market ratio, stock variance, long-term yield, long-term return, term spread, default return spread, default return yield, inflation and net equity expansion. A detailed description, including data sources, can be found in Appendix B.[4] All of these variables have been proposed in the literature as having some degree of predictability, in-sample. Appendix C verifies these findings. However, the vital question is whether this predictability exists out-of-sample and whether an investor can benefit from it in real time. In order to investigate this further we construct eleven single predictor forecasting models, an unrestricted kitchen sink model, five combination forecast models and a kitchen sink model estimated via the James-Stein estimator. As mentioned in the introduction, we will use the historical average as a benchmark model:

$\hat{\mu}_{t+1|t}^{Benchmark} = \frac{1}{R}\sum_{j=t-R+1}^{t} Y_j.$

As already discussed, we will use a rolling estimation window to estimate the parameters.

[4] A special thanks to Amit Goyal for borrowing his dataset.

The length of the estimation window is $R = 360$ (i.e. 30 years) and this implies that $\hat{\pi} = P/R = 1.8$ if we consider the entire forecast evaluation period. We consider three different out-of-sample forecast evaluation periods: 1957:1-2010:12, 1957:1-1983:12 and 1984:1-2010:12. Overall, the consideration of multiple out-of-sample periods helps to provide us with a good sense of the robustness of the out-of-sample forecasting results.

In addition to the $OOS - t$ test statistic mentioned in the previous section, we will also compute the measure $R^2_{OOS} = 1 - \frac{\sum_{t=R}^{R+P}(y_t - \hat{\mu}_{t+1|t}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}))^2}{\sum_{t=R}^{R+P}(y_t - \hat{\mu}^{Benchmark}_{t+1|t})^2}$. This measure is positive if model j has a lower root mean squared error than the historical average and is measured in the same units as the in-sample R^2. Table 3.1 illustrates the forecasting performance of selected forecasting models relative to the constant mean model as measured by their $OOS - t$ test statistic and R^2_{OOS}.

Generally, the single predictor forecasting models performs poorly but the *models* based on inflation and the long term government bond return outperform the historical average at a 5% level when evaluated over the entire time-period [5]. However, none of the individual forecasting methods outperform the simple average if we penalise the forecasts for estimation error. This is interesting from a practical point of view. We reject that the expected loss of the true forecasting model based on inflation or the default return spread is equal to the expected loss from the benchmark model. However, it will never be feasible for a forecaster to infer the true model since the data sample is finite. This is particularly true if there are structural breaks over the time period such that it is not possible to utilise the entire sample to estimate the present relationship. The end-user of the forecast is then forced to consume not only the true forecast value, but the estimation error as well:

$$\hat{\mu}^i_{t+1|t}(\hat{\boldsymbol{\beta}}) = \hat{\mathbb{E}}[Y_{t+1}] + \boldsymbol{\beta}_0^T \tilde{\mathbf{X}}_t + \underbrace{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \tilde{\mathbf{X}}_t}_{Estimation\,Error}.$$

Therefore, from a practical point of view, none of the individual forecasting methods outperform the historical average significantly. The multivariate kithchen sink model also

[5] Note that the ARMA model has lower critical values than the other models since it has 3 parameters rather than 2

	1957:1-2010:12		1957:1-1983:12		1984:1-2010:12	
	$\hat{\pi} = P/R \simeq 1.8$		$\hat{\pi} = P/R \simeq 0.9$		$\hat{\pi} = P/R \simeq 0.9$	
Forecasting Model	$OOS-t$	R^2_{OOS}	$OOS-t$	R^2_{OOS}	$OOS-t$	R^2_{OOS}
Single Predictor Models:						
Dividend-price ratio	0.02	0.02	$0.71^{\star\star}$	1.18	-1.1	-0.9
Earnings-price ratio	-1.07	-1.01	-0.73	-1.02	-0.8	-1.0
Book-to-market ratio	-1.31	-1.21	-0.97	-1.98	-2.0	-0.6
Stock variance	0.02	0.02	-1.78	-1.16	$0.6^{\star}$	1.0
LT government bond yield	-0.73	-1.19	-0.54	-1.93	-1.6	-0.6
LT government bond return	$0.52^{\star\star}$	0.51	$0.93^{\star\star}$	1.64	-0.4	-0.4
Term spread	0.02	0.03	$0.65^{\star}$	1.66	-1.0	-1.3
Default return spread	-0.89	-0.97	-1.00	-1.05	-0.5	-0.9
Default return yield	-1.39	-0.96	-0.46	-0.51	-1.5	-1.3
Inflation	$0.71^{\star\star}$	0.74	$1.46^{\star\star\star}$	2.00	-0.2	-0.3
Net equity expansion	-0.32	-0.27	$0.51^{\star}$	0.68	-1.0	-1.0
Combination Forecasts:						
Mean	$2.54^{\star\star\star}_{\blacktriangle\blacktriangle\blacktriangle}$	1.02	$2.77^{\star\star\star}_{\blacktriangle\blacktriangle\blacktriangle}$	1.92	$0.71^{\star\star}$	0.31
Trimmed mean	$2.36^{\star\star\star}_{\blacktriangle\blacktriangle\blacktriangle}$	0.85	$2.72^{\star\star\star}_{\blacktriangle\blacktriangle\blacktriangle}$	1.81	0.25	0.08
Median	$1.58^{\star\star}_{\blacktriangle}$	0.44	$2.06^{\star\star\star}_{\blacktriangle\blacktriangle}$	1.12	-0.57	-0.11
Optimal weights	-1.60	-3.83	-0.04	-0.10	-1.02	-6.75
Opt. W. (No correlation)	$2.54^{\star\star\star}_{\blacktriangle\blacktriangle\blacktriangle}$	1.02	$2.77^{\star\star\star}_{\blacktriangle\blacktriangle\blacktriangle}$	1.92	$0.71^{\star}$	0.31
Kitchen Sink Model:						
OLS estimator	-1.03	-4.68	-0.14	-0.43	-1.05	-8.08
James-Stein estimator	$2.07^{\star\star\star}_{\blacktriangle\blacktriangle\blacktriangle}$	1.69	$2.51^{\star\star\star}_{\blacktriangle\blacktriangle\blacktriangle}$	2.64	$0.80^{\star\star}$	1.26

Table 3.1: Out-of-Sample Equity Premium Forecasting Results

Notes: The table presents statistics on relative forecasting performance for CRSP excess returns at a 1-month forecast horizon. The benchmark prediction model is the historical average. Estimation is based on a rolling estimation scheme with a window size of $R = 360$. R^2_{OOS} is an out-of-sample R^2 statistic. $OOS-t$ is a test statistic used to test if a *model* or a *method* has the same expected loss as the historical average. Significant evidence against the null that the two *models* have the same expected loss is indicated by $\bigstar's$: $\bigstar$ 10%, $\bigstar\bigstar$ 5% and $\bigstar\bigstar\bigstar$ 1% level respectively. Significant evidence against the null that the two *methods* have the same expected loss is indicated by $\blacktriangle's$: $\blacktriangle$ 10%, $\blacktriangle\blacktriangle$ 5% and $\blacktriangle\blacktriangle\blacktriangle$ 1% level respectively.

performed poorly, which is consistent with the empirical results reported in Welch & Goyal (2008).

Interestingly, combinations of individual forecasts generally perform better than the individual forecast themselves, which is consistent with the empirical results reported in Rapach, Strauss & Zhou (2010). Four out of five forecast combination methods outperform the historical average at a 1% significance level. Even if we penalise the calculations for estimation errors these combination methods outperform the historical average. There are, therefore, substantial statistical gains to be made from combining individual single predictor forecasts.

Another interesting finding is that the kitchen sink model, estimated by the James-Stein method, performs substantially better than the unrestricted model (estimated via ordinary least squares). This approach also significantly outperforms the constant mean model. Furthermore, the out-of-sample predictive power, as measured by the R^2_{OOS}, is larger than any of the combination forecast methods. As we showed in the previous section, the combination forecast approach is effectively a multivariate, kitchen sink model with shrinked model parameters. However, it remains unclear whether the model parameters are 'shrinked' in an optimal manner . We will analyse this issue further in the following subsection.

3.4.1 An Equal Weights Shrinkage Decomposition

We will focus our attention on the static mean combination forecast due to its simplicity. Based on equation 3.3.5, it is clear that the (equal weights) mean combination forecast is given by:

$$\hat{\mu}^{MCF}_{t+1|t}(\hat{\boldsymbol{\beta}}_{SPE,t}, \frac{N-1}{N}\iota) = \hat{\mu}^{KS}_{t+1|t}(\hat{\boldsymbol{\beta}}_{OLS,t}, \mathbf{0}) + \underbrace{N^{-1}\sum_{i=1}^{k}\beta^{OLS}_{i,t}\tilde{X}_{i,t}}_{Shrinkage} + \underbrace{N^{-1}\sum_{i=1}^{k}(\hat{\beta}^{SPE}_{i,t} - \hat{\beta}^{OLS}_{i,t})\tilde{X}_{i,t}}_{Bias\,Correction}.$$

The mean combination forecast outperformed the constant mean model while the kitchen sink model performed very poorly. Based on the decomposition above, a natural question is whether the strong performance of the mean combination forecast is due to 'shrinkage' or to

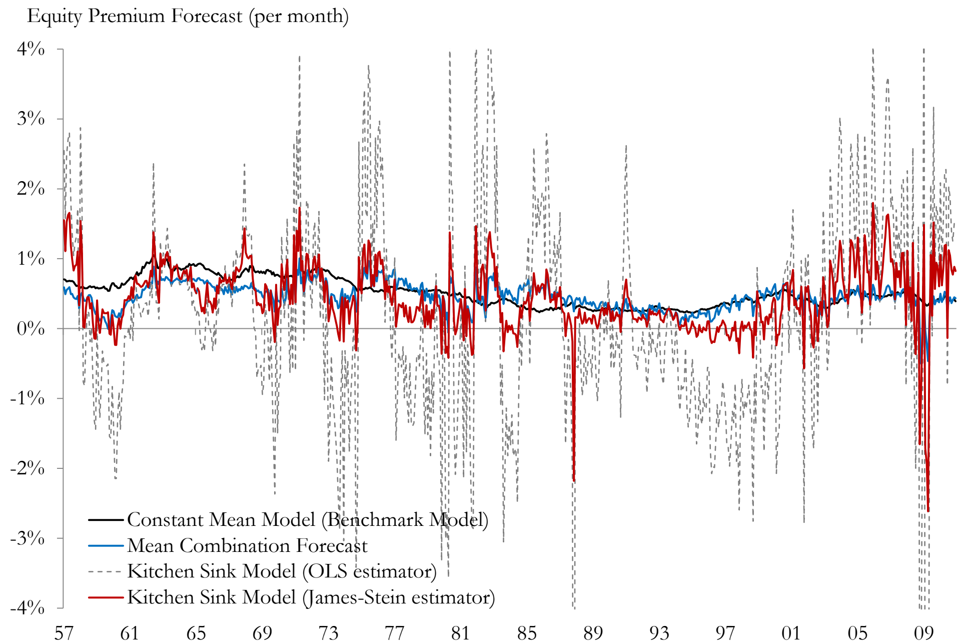


Figure 3.3: *Monthly equity premium forecasts over the period 1957:1-2010:12*

the 'bias correction'. In order to evaluate this in more detail, we analyse the performance of the two types of models for different (common) shrinkage levels:

$$\hat{\mu}_{t+1|t}^{OLS(\lambda)}(\hat{\boldsymbol{\beta}}_{OLS,t}, \lambda\boldsymbol{\iota}) = \hat{\mathbb{E}}[Y_{t+1}] + (1-\lambda)\sum_{i=1}^{k}\hat{\beta}_{i,t}^{OLS}(X_{i,t} - \hat{\mathbb{E}}[X_{i,t}])$$

and

$$\hat{\mu}_{t+1|t}^{SPE(\lambda)}(\hat{\boldsymbol{\beta}}_{SPE,t}, \lambda\boldsymbol{\iota}) = \hat{\mathbb{E}}[Y_{t+1}] + (1-\lambda)\sum_{i=1}^{k}\hat{\beta}_{i,t}^{SPE}(X_{i,t} - \hat{\mathbb{E}}[X_{i,t}]).$$

Clearly, these two models nest both the kitchen sink and the mean combination forecast model: for $\lambda = 0$, $\hat{\mu}_{t+1|t}^{OLS(\lambda)}(\hat{\boldsymbol{\beta}}_{OLS,t}, \lambda\boldsymbol{\iota})$ is the unrestricted kitchen sink model and for $\lambda = \frac{N-1}{N}$ we have that $\hat{\mu}_{t+1|t}^{SPE(\lambda)}(\hat{\boldsymbol{\beta}}_{SPE,t}, \lambda\boldsymbol{\iota})$ is the mean combination forecast. Figure 3.4 illustrates the R_{OOS}^2 for the different degrees of shrinkage for each of the two models.

The figure illustrates clearly that the kitchen sink model performs poorly while the mean combination forecast approach outperforms the constant mean model. However, a more interesting observation is that the mean combination forecast appears to 'overshrink' the model parameters. The figure also shows that the James-Stein estimator performs

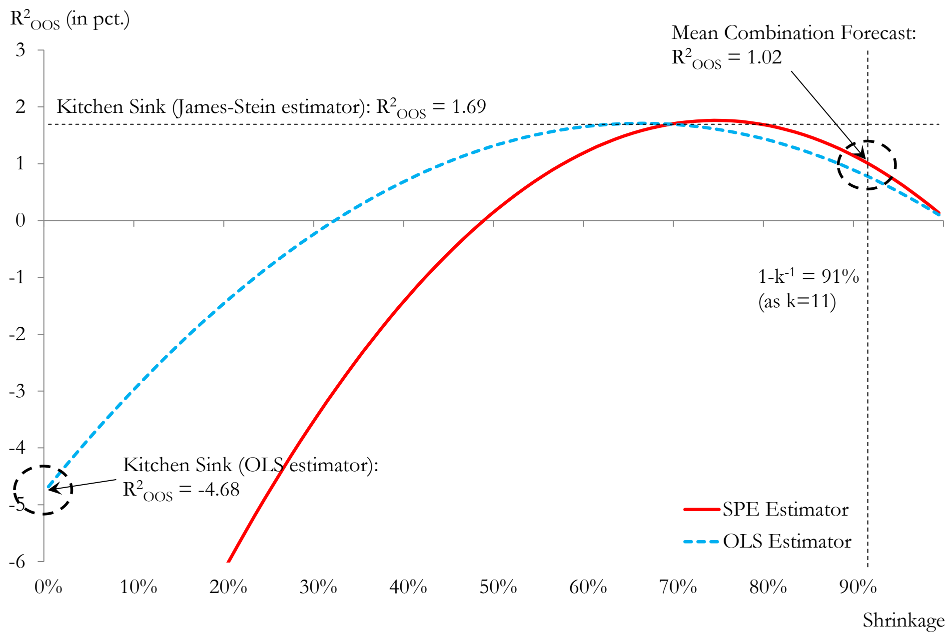


Figure 3.4: *Illustration of equal weights shrinkage decomposition:* R^2_{OOS} *of* $\hat{\mu}_{t+1|t}(\boldsymbol{\beta}, \lambda\boldsymbol{\iota})$ *based on* $\lambda \in [0:1]$ *and* $\boldsymbol{\beta} \in \{\hat{\boldsymbol{\beta}}_{SPE,t}, \hat{\boldsymbol{\beta}}_{OLS,t}\}$.

substantially better than the mean combination forecasts. Figure 3.5 illustrates that the average shrinkage in the James-Stein estimator is $75\% < 91\% = \frac{N-1}{N}$. Therefore, the empirical evidence suggests that the combination forecasts from single predictor models tend implicitly to overshrink the model parameters. Figure 3.3 illustrates the equity premium forecasts based on the unrestricted kitchen sink model, the mean combination forecast and the James-Stein kitchen sink model. This figure clearly illustrates that the unrestricted kitchen sink model tends to generate very noisy and volatile estimates of the unobserved equity premium. However, the results in this section suggest that the mean combination forecast tends to overshrink the model parameters. Thus, a James-Stein type estimator appears to be superior. Finally, it remains unclear whether the bias correction term improves the model performance or not. If one has a quick look at figure 3.4 the two estimators appear to perform equally well, although the optimal amount of shrinkage is higher with the SPE estimator.

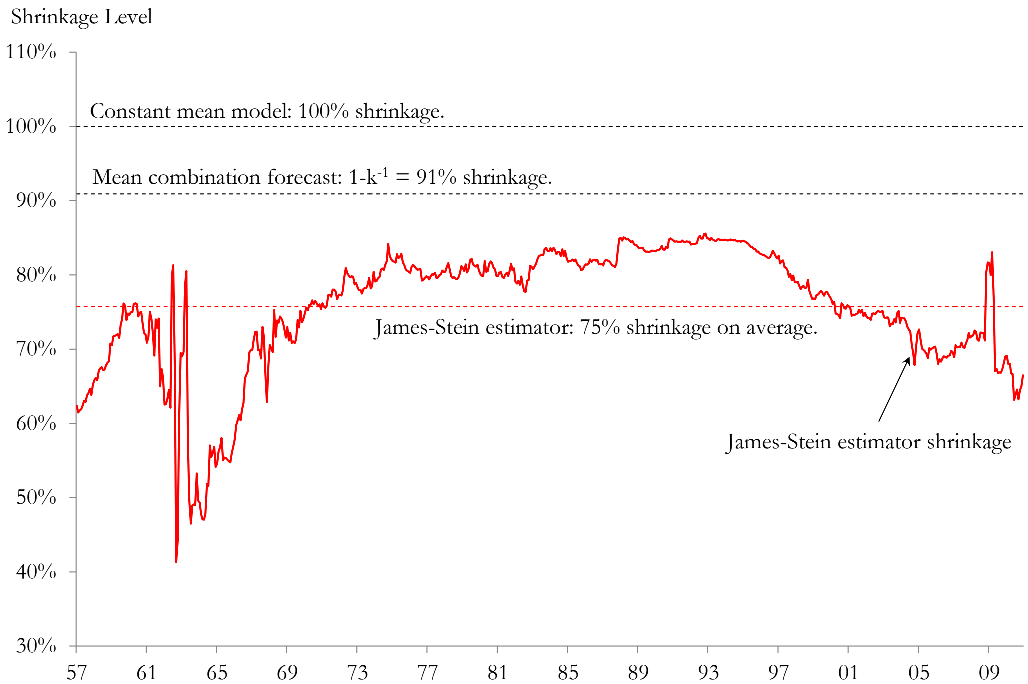


Figure 3.5: *Illustration of James-Stein estimator shrinkage over the period 1957:1-2010:12 relative to the mean combination forecast (91% shrinkage) and the constant mean model (100% shrinkage).*

3.5 Estimating the Economic Value of Predictability

In the previous section we showed that a kitchen sink model, estimated by a James-Stein estimator, and combination forecast methods outperform a constant mean model significantly. However, the R^2_{OOS} of these methods were only around 1-2% . This raises the question of whether an investor in real time could benefit from the predictability. To investigate this issue further we revisit our parsimonious model for an investor's portfolio decision.

We will again assume that we are in a two-period world with only two assets: a riskless bond that yields a safe gross return of R_0 and a risky asset with a stochastic gross return of R. The risky asset can be interpreted as the aggregate stock market index. The investor chooses the fraction of his wealth to be invested in the risky asset at time t, α_t. His wealth at period

$t+1$ is stochastic and a function of α_t: $W_{t+1}(\alpha_t) = W_t(R_0 + \alpha_t\{R_{t+1} - R_0\})$. Under assumptions 1-3, it is easy to verify that the investor's optimal portfolio is given by

$$\alpha_t^*(\mu_{t+1|t}, \sigma^2; \gamma) = \frac{\mu_{t+1|t}}{\gamma\sigma^2},$$

where $\mu_{t+1|t}$ denotes the conditional mean, σ^2 denotes the homoskedastic variance and γ denotes the investor's risk aversion. Intuitively, the fraction invested in the risky asset depends positively on the conditional mean and negatively on the conditional volatility and risk aversion. However, a real-world investor does not know the true conditional mean, $\mu_{t+1,t}$, and variance, $\sigma^2_{t+1,t}$, and therefore it is not feasible to determine α_t^*. He will have to rely on the estimates $\hat{\mu}_{t+1|t}$ and $\hat{\sigma}^2$ in order to determine his optimal portfolio. Generally, more precise estimates of the conditional mean and variance should lead to more efficient investment decisions for the investor. We will use our results from the previous section as forecasts of the equity premium. However, in order to implement the trading strategy the investor will also need to estimate the variance. We will simply use the sample variance

$$\hat{\sigma}^2_{\tau+1|\tau} = \frac{1}{R}\sum_{t=\tau-R+1}^{\tau}(Y_t - \overline{Y})^2,\ \tau = R, ..., R+P-1,$$

such that the investor's optimal portfolio is: $\alpha_t^i = \frac{\hat{\mu}^i_{t+1|t}}{\gamma\hat{\sigma}^2_{t+1|t}}$, $t = R, ..., R+P-1$. In order to evaluate the utility gained from using more accurate forecasts of the conditional mean, we will use the P out-of-sample observations to estimate the investor's expected utility. As illustrated in Appendix A, it is possible to rewrite the investor's expected utility as $\mathbb{E}[u(w_{t+1})|\mathcal{F}_t] = \mathbb{E}[w_{t+1}|\mathcal{F}_t] - \frac{\gamma}{2}\mathbb{V}[w_{t+1}|\mathcal{F}_t]$. Therefore, we estimate the expected utility as

$$\hat{\mathbb{E}}[u(w_{t+1}(\hat{\alpha}^*_{t,i}))|\mathcal{F}_t] = \frac{1}{P}\sum_{t=R}^{R+P-1} W_{t+1}(\alpha_t^i) - \frac{\gamma}{2}\frac{1}{P-1}\sum_{t=R}^{R+P-1}\left(W_{t+1}(\alpha_t^i) - \bar{W}(\alpha_t^i)\right)^2.$$

Since the marginal utility of expected return is unity, the difference in expected utility between two investment strategies is equal to the certainty equivalent return. Therefore, the difference in expected utility can be interpreted as the portfolio management fee an investor is willing to pay to have access to the additional information available in a given forecast. In addition to the historical average, we will also consider three simple static strategies: (i)

$\alpha_t = 0$, $\forall t$, (ii) $\alpha_t = \frac{1}{2}$, $\forall t$ and (iii) $\alpha_t = 1$, $\forall t$. Table 3.2 shows the estimated expected return, volatility and expected utility from the investment strategies based on the four dominating combination forecasts, the kitchen sink model (using OLS or the James-Stein estimator) and the benchmark forecasts. We consider both the case where no short sales or borrowing is allowed, $\alpha_t \in [0:1]$, and the case where there are no restrictions, $\alpha_t \in \mathbb{R}$.

It is clear from table 3.2 that the predictability in the equity premium is of economic value to a investor despite a relative low R^2_{OOS}. A CARA investor, with $\gamma = 3$ and who bases his investment decision on the mean-combination forecast, will gain around 80-90 basis points of certainty equivalent return annually depending on whether short sales and borrowing are allowed or not. This corresponds to a non-trivial increase in expected utility from around 3.3% to 4.1 or 4.2%. For the James-Stein approach the gains are even larger. The investor will gain around 150-160 basis points of certainty equivalent returns. By using more accurate forecasts of the conditional mean the investor obtains both a higher expected return and lower volatility on his portfolio.

3.6 Analysing the Economic Gains from Weak Predictability

In the empirical section we showed that the equity premium is predictable once one imposes certain restrictions on the model parameters. However, the R^2_{OOS} were only around 1-2% and this raised the question of whether the equity premium predictability is economically meaningful for an investor. In the previous section we attempted to answer this question by estimating the gains in expected utility that would be derived from using more accurate forecasts of the equity premium. Here, we attempt to answer the same question in a more analytical way.

Recall the setup from 3.3 where we proposed the following general model for excess returns: $Y_{t+1} = \mathbb{E}[Y_{t+1}|\mathcal{F}_t] + \epsilon_{t+1}$, where $\mathbb{E}[Y_{t+1}|\mathcal{F}_t]$ denotes the equity premium and ϵ_{t+1} is an unobserved error that has a mean of zero and a constant variance of σ^2. Furthermore, when deriving the investor's optimal portfolio decision, we also assumed that the risky asset had a

A: Short sales and borrowing allowed: $\alpha_t \in \mathbb{R}$

	$\widehat{\mathbb{E}[W]}$	$\widehat{\mathbb{V}[W]}^{0.5}$	$\mathbb{E}[\widehat{U(W)}]$	Utility Gain
Combination Forecasts:				
Mean	4.7%	6.5%	4.1%	0.83%
Trimmed Mean	4.5%	6.5%	3.9%	0.64%
Median	4.2%	6.8%	3.5%	0.27%
Optimal Weights	5.1%	20.2%	-1.1%	-4.35%
No Correlation	4.7%	6.5%	4.1%	0.83%
Kitchen Sink Model:				
James-Stein Estimator	6.1%	9.3%	4.8%	1.57%
OLS estimator	12.3%	38.1%	-9.4%	-12.66%
Benchmark Forecasts:				
All bonds	2.9%	0.8%	2.9%	-
50% Bonds and 50% Equity	4.9%	7.5%	4.1%	-
All Equity	7.0%	15.0%	3.6%	-
Constant Mean Model	4.0%	7.0%	3.3%	-

B: No short sales or borrowing: $\alpha_t \in [0 : 1]$

	$\widehat{\mathbb{E}[W]}$	$\widehat{\mathbb{V}[W]}^{0.5}$	$\mathbb{E}[\widehat{U(W)}]$	Utility Gain
Combination Forecasts				
Mean	4.8%	6.5%	4.2%	0.89%
Trimmed Mean	4.6%	6.5%	4.0%	0.73%
Median	4.2%	6.8%	3.5%	0.27%
Optimal Weights	5.5%	9.8%	4.0%	0.75%
No Correlation	4.8%	6.5%	4.2%	0.89%
Kitchen Sink Model:				
James-Stein Estimator	5.5%	7.0%	4.8%	1.52%
OLS estimator	6.7%	8.1%	5.7%	2.43%
Benchmark Forecasts:				
All bonds	2.9%	0.8%	2.9%	-
50% Bonds and 50% Equity	4.9%	7.5%	4.1%	-
All Equity	7.0%	15.0%	3.6%	-
Constant Mean Model	4.0%	7.0%	3.3%	-

Table 3.2: The Economic Value of Predictability.

Notes: The table presents statistics on portfolio performance when the investment decision is based on different forecasts of the equity premium. All values are annualised. The investor has CARA preferences with $\gamma = 3$. The utility gains are relative to the constant mean model. Estimation is based on a rolling estimation window with size $R = 360$ and the evaluation period is 1957:1-2010:12.

Gaussian distribution. Here, we attempt to measure how much value a $R^2_{OOS} = 1\%$ gives to an investor. Following Campbell & Thompson (2008), we simply assume that $\mathbb{E}[Y_{t+1}|\mathcal{F}_t] = \mu + \beta x_t$ where μ is the unconditional equity premium and x_t is a predictor variable with a mean of zero and a constant variance of σ_x^2. Following the setup of proposition 1, we consider an investor with a single-period horizon and CARA preferences. If the investor does not observe x_t, the conditional mean and variance is given by μ and $\sigma^2 + \sigma_\epsilon^2$. On the other hand, if the investor does observe x_t, the conditional mean and variance is given by $\mu + \beta x_t$ and σ^2 respectively. Therefore, the optimal portfolio is given by $\alpha = \gamma^{-1}\mu(\sigma^2 + \sigma_x^2)^{-1}$ if the investor does not observe x_t but $\alpha_t^* = \gamma^{-1}(\mu + \beta x_t)\sigma^{-2}$ if he does observe x_t. Consequently, if the investor does not observe x_t, the expected utility is given by

$$\begin{aligned}
\mathbb{E}[u(w_{t+1}(\alpha))|\mathcal{F}_t] &= \frac{\mu^2}{\gamma(\sigma^2 + \beta^2\sigma_x^2} - \frac{1}{2\gamma}\frac{\mu^2\sigma^2}{(\sigma^2 + \beta^2\sigma_x^2)^2} \\
&= \left(\frac{1}{2\gamma}\right) S^2(1 + R^2)\,,
\end{aligned}$$

where $S = \frac{\mu}{\sigma^2 + \beta^2\sigma_x^2}$ is the unconditional Sharpe ratio, and $R^2 = \frac{\beta^2\sigma_x^2}{\beta^2\sigma_x^2 + \sigma^2}$, where R^2 is the statistic for the regression of excess return on the predictor variable, x_t. On the other hand, if the investor observes x_t the expected utility is given by

$$\begin{aligned}
\mathbb{E}[u(w_{t+1}(\alpha_t^*))|\mathcal{F}_t] &= \left(\frac{1}{2\gamma}\right)\left(\frac{\mu + \beta^2\sigma_x^2}{\sigma^2}\right). \\
&= \left(\frac{1}{2\gamma}\right)\left(\frac{S^2 + R^2}{1 - R^2}\right)
\end{aligned}$$

The difference in expected utility is given by

$$\left(\frac{R^2}{2\gamma}\right)\left(\frac{1 + S^2R^2}{1 - R^2}\right) \geq \left(\frac{R^2}{2\gamma}\right),$$

which is always larger than $R^2/2\gamma$ and is close to $R^2/2\gamma$ when the time interval is short such that R^2_{OOS} and S^2 are both small. In terms of the proportional increase in expected utility

$$\frac{\mathbb{E}[u(w_{t+1}(\alpha_t^*))|\mathcal{F}_t] - \mathbb{E}[u(w_{t+1}(\alpha))|\mathcal{F}_t]}{\mathbb{E}[u(w_{t+1}(\alpha))|\mathcal{F}_t]} = \left(\frac{R^2}{S^2}\right)\left(\frac{1 + S^2R^2}{1 - R^4}\right) \geq \left(\frac{R^2}{S^2}\right),$$

which is always larger than R^2/S^2 and close to R^2/S^2 when the time interval is short and R^2 and S^2 are both small. This analysis shows that the correct way to judge the magnitude of

R^2_{OOS} is to compare it with the squared (unconditional) Sharpe ratio. Campbell & Thompson (2008) considered a similar model framework and found that the proportional increase in expected return was given by

$$\frac{\mathbb{E}[w_{t+1}(\alpha_t^*)|\mathcal{F}_t] - \mathbb{E}[w_{t+1}(\alpha)|\mathcal{F}_t]}{\mathbb{E}[w_{t+1}(\alpha)|\mathcal{F}_t]} = \left(\frac{R^2}{S^2}\right)\left(\frac{1+S^2}{1-R^2}\right).$$

However, the investor who observes x_t gets a higher portfolio return in part by taking on greater risk. Thus, the increase in the average return is not a pure welfare gain for a risk-averse investor as $\left(\frac{1+S^2R^2}{1-R^4}\right) < \left(\frac{1+S^2}{1-R^2}\right)$. This was also acknowledged by Campbell and Thompson.

Over the period 1957:1-2010:12 the (annualized) equity premium was 7% and the (annualized) volatility was 15%. This translates into a monthly Sharpe ratio of 13.5% and a squared Sharpe ratio of 1.81%. Importantly, based on the formula above, a R^2 of 1% translates into a 55% increase in expected utility (from 3.7% to 5.7%) and 57% increase in expected return. In comparison, for the James-Stein approach, we found that the improvement was 48% (from 3.3% to 4.8%).

3.7 Conclusion

In this paper we have reassesed the statistical evidence of equity premium predictability. Generally, we cannot reject the hypothesis that all the single predictor methods have the same expected loss as a constant mean model. This confirms the findings in Welch & Goyal (2008).

We found that methods used to combine individual forecasts performed well; and even if we penalise the calculations for estimation errors we cannot reject the assertion that they have the same expected loss as the historical average at a 1% level. However, an even more promising approach appears to be based on the kitchen sink model when estimated via a James-Stein type estimator. The combination forecast approach tended to implicitly overshrink the model parameters. The success of the James-Stein approach is interesting from a practical point of view. It is not feasible to use the true forecasting model for a real-life investor because he is forced to rely on estimated parameter values.

In order to estimate the value of this out-of-sample equity premium predictability, we specified the preferences of a representative investor and estimated his optimal portfolio choice. Based on his risk aversion as well as whether short sales and borrowing were allowed, a mean-variance investor is willing to pay around 80-90 basis points of certainty equivalent return in order to obtain the information from the combination forecasts; and he is willing to pay 150-160 basis points when calculations are based on the James-Stein approach.

The case for stock market predictability seems stronger than the conclusions in Welch & Goyal (2008) lead one to believe. However, it should be emphasised that we have not taken trading costs into account. In addition, it is not free to have a stock market analyst to perform this analysis. After we have corrected for trading costs and a realistic management fee there might not be any predictability left.

3.8 Appendix

Appendix A: Proof of Proposition 1

We assume that we are in a two-period world with only two assets: a riskless bond that yields a safe gross return of R_0 and a risky asset with a stochastic excess return of R_{t+1}. Furthermore, we will make the following assumptions:

A.1 (Utility Maximization) *The investor chooses his optimal portfolio, α_t^*, so as to maximize his expected utility in the second period,*

$$\alpha_t^* = \underset{\alpha_t \in \mathbb{R}}{arg\,max}\, \mathbb{E}[u(w_{t+1}(\alpha_t);\gamma)|\mathcal{F}_t]\,,$$

where $u(\cdot)$ denotes the investor's utility function, $w_{t+1}(\alpha_t) = \alpha_t R_{t+1} + (1-\alpha_t)R_0$ denotes the investor's wealth in the second period, γ is a parameter that determines the investor's risk aversion and $\mathcal{F}_t$ is a $\sigma - algebra$ that represents the investor's information set in the initial period, t.

A.2 (Distribution of Risky Asset) *The risky return has the following conditional distribution:*

$$R_{t+1}|\mathcal{F}_t \sim \mathcal{N}(R_0 + \mu_{t+1|t}, \sigma^2),$$

where $\mu_{t+1|t}$ denotes the conditional equity premium, and σ^2 denotes the variance.

A.3 (Investor Preferences) *The investor's preferences can be represented by a constant absolute risk aversion (CARA) utility function*

$$u(w_{t+1};\gamma) = -exp(-\gamma w_{t+1}),$$

where γ describes the degree of risk aversion.

Proposition 1 (Investor Loss Function): *Suppose that assumptions A.1-A.3 above hold. Then, the investor's loss associated with forecasting the conditional mean with error can be* *described by the following loss function:*

$$\begin{aligned}\mathcal{L}(\mu_{t+1|t}, \hat{\mu}_{t+1|t}; \gamma, \sigma^2) &= \mathbb{E}[u(w_{t+1}(\alpha_t(\mu)))|\mathcal{F}_t] - \mathbb{E}[u(w_{t+1}(\alpha_t(\hat{\mu})))|\mathcal{F}_t] \\ &= \frac{(\mu_{t+1|t} - \hat{\mu}_{t+1|t})^2}{2\gamma\sigma^2}.\end{aligned}$$

Similarly, the representative investor's loss associated with forecasting the variance with error can be described by the following loss function:

$$\begin{aligned}\mathcal{L}(\sigma^2, \hat{\sigma}^2; \gamma, \mu_{t+1|t}) &= \mathbb{E}[u(w_{t+1}(\alpha_t(\sigma^2)))|\mathcal{F}_t] - \mathbb{E}[u(w_{t+1}(\alpha_t(\hat{\sigma}^2)))|\mathcal{F}_t] \\ &= \mu_{t+1|t}^2 \gamma^{-1} \left\{ \frac{1}{2\sigma^2} + \frac{\sigma^2}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^2} \right\}.\end{aligned}$$

Proof: First, we solve for the investor's optimal portfolio decision, α^*, in order to evaluate how this depends on the conditional mean and variance of the risky asset. In A.2 and A.3 the investor chooses his portfolio in order to maximize wealth in period $t+1$ such that

$$\alpha_t^* = \underset{\alpha_t \in \mathbb{R}}{arg\,max}\, \mathbb{E}[-exp(-\gamma w_{t+1}(\alpha_t)|\mathcal{F}_t].$$

By A.1, it is clear that the investor's wealth in the second period is normally distributed: $w_{t+1}(\alpha_t)|\mathcal{F}_t \sim \mathcal{N}(\alpha_t \mu_{t+1|t} + (1-\alpha_t)R_o, \alpha_t^2 \sigma_t)$. Importantly, this implies that the utility of wealth in period $t+1$ is log-normally distributed,
$exp(-\gamma w_{t+1}(\alpha_t))|\mathcal{F}_t \sim log\mathcal{N}(-\gamma\{\alpha_t \mu_{t+1|t} + (1-\alpha_t)R_o\}, \gamma^2 \alpha_t^2 \sigma^2)$, and we can therefore rewrite the investor's optimization problem as

$$\alpha_t^* = \underset{\alpha_t \in \mathbb{R}}{arg\,max} \left\{ -exp(-\gamma\{\alpha_t \mu_{t+1|t} + (1-\alpha_t)R_o\} + \frac{1}{2}\gamma^2 \alpha_t^2 \sigma^2) \right\}.$$

Furthermore, after a few positive monotone transformations, we can rewrite the objective function as $\mathbb{E}[w_{t+1}(\alpha_t)|\mathcal{F}_t] - \frac{\gamma}{2}\mathbb{V}[w_{t+1}(\alpha_t)|\mathcal{F}_t]$, and we can then easily solve for the optimal portfolio:

$$\alpha_t^*(\mu_{t+t|t}, \sigma^2; \gamma) = \frac{\mu_{t+1|t} - R_0}{\gamma\sigma^2}.$$

Intuitively, the investor will allocate a larger fraction of his wealth to the risky asset if the equity premium is higher, if the volatility is lower or if the investor's risk aversion decreases. Given the equity premium and the volatility of the risky asset, the investor's optimal portfolio will lead to an expected utility given by $\mathbb{E}[u(w_{t+1}(\alpha_t))|\mathcal{F}_t] = \mu_{t+1|t}^2/(2\gamma\sigma^2)$.

Unfortunately for the representative investor, the equity premium and volatility of the risky premium is unobservable. In order to evaluate how costly it is for the investor to forecast the conditional mean and the volatility with error, we derive the investor's loss function, defined over volatility or conditional mean forecast errors, as the loss in expected utility:
$\mathbb{E}[u(w_{t+1}(\alpha_t(\hat{\mu}_{t+1|t}))|\mathcal{F}_t] = (\mu_{t+1|t}\hat{\mu}_{t+1|t} - \frac{1}{2}\mu^2_{t+1|t}) \cdot (\gamma\sigma^2)^{-1}$ and
$\mathbb{E}[u(w_{t+1}(\alpha_t(\hat{\sigma}^2)))|\mathcal{F}_t] = \mu^2_{t+1|t} \cdot (\gamma\hat{\sigma}^2)^{-1} - (\mu^2_{t+1|t}\sigma^2) \cdot (2\gamma\hat{\sigma}^4)^{-1}$. Proposition 1 follows. $\square$

Appendix B: Data Description

The data frequency is monthly and covers the period 1927:1-2010:12. The explanatory variables are from Welch and Goyal (2008) and the equity premium is from Kenneth French.

- **The Equity Premium:** The total rate of return on the stock market minus the short-term interest rate. As stock returns, we use the value-weighted CRSP index returns, including dividends, from Center for Research in Security Prices (CRSP). The index is based on securities traded on NYSE, NYSE Amex, and NASDAQ. The returns are continuously compounded, $r_t = lnP_t - lnP_{t-1}$. The risk-free rate is the Treasury-bill rate.
- **Dividend-Price Ratio (D/P):** Difference between the log of dividends and the log of prices. Dividends are 12-month moving sums of dividends paid on the S&P500 index. Dividends are from Robert Shiller's website for 1926-1987 and from S&P Corporation for 1988-2008. Prices are the level of the S&P composite index from CRSP.
- **Earnings-Price Ratio (E/P)**: Difference between the log of earnings and the log of prices. Earnings are 12-month moving sums of earnings on the S&P 500 index. From 1926-1987 earnings are from Robert Shiller's website and for 1988-2008 they are based on interpolations of quarterly earnings provided by the S&P Corporation.
- **Book-to-Market Ratio (B/M)**: The ratio of book value to market value for the Dow Jones Industrial Average. For the months from March to December, this is computed by dividing book value at the end of the previous year by the price at the end of the current month. For the months of January and February, this is computed by dividing book value at the end of two years ago by the price at the end of the current month. Book values are from Value Line's website, specifically their Long-Term Perspective Chart of the Dow Jones Industrial average. **M**arket values are for the Dow Jones industrial average.
- **Stock Variance (SVAR)**: The stock variance is unobservable but we estimate the monthly variance by the sum of squared daily returns on the S&P 500 index.

- **Long Term Yield (LTY)**: Long-term government bond yields are from Ibbotson's Stocks, Bonds, Bills and Inflation Yearbook.
- **Long Term Return (LTR)**: Long-term government bond returns are from Ibbotson's Stocks, Bonds, Bills and Inflation Yearbook.
- **Term Spread (TMS)**: Difference between the long-term goverment yield and the Treasury bill rate. Treasury bill rates from 1926-1933 are from US Yields on Short-Term United States Securities, Three-Six Month Treasury Notes and Certificates in the NBER Macroeconomic data base. For 1934-2009 we use the 3-Month Treasury Bill: Secondary Market Rate from the economic research data base at the Federal Reserve Bank at St. Louis (FRED)
- **Default Return Spread (DFR)**: Difference between the long-term corporate bonds and long-term government bond returns. Long-term government bonds are from Ibbotson's Stocks, Bonds, Bills and Inflation Yearbook.
- **Default Return Yield (DFY)**: Difference between BAA and AAA-rated corporate bond yields. The bond yields are from FRED.
- **Inflation (INFL)**: Calculated from the CPI from the Bureau of Labor Statistics. Because inflation information is released only in the following month we lag the variable one month.
- **Net Equity Expansion (NTIS)**: It is calculated as the ratio of 12 month moving sums of net issues by NYSE listed stocks divided by the total end-of-year market capitalisation of NYSE stocks. The dollar amount of net equity issuing activity (IPO's, SEO's, stock repurchases, less dividends) for NYSE listed stocks is computed from CRSP data as: $Net\,Issue_t = Mcap_t - Mcap_{t-1} \cdot (1 + r_t^{vw,nd})$ where $Mcap$ is the total market capitalisation, and $r_t^{vw,nd}$ is the value weighted return (excluding dividends) on the NYSE index.

For further details see Amit Goyal's website (http://www.bus.emory.edu/agoyal/Research.html). Figure 3.6 and 3.7 illustrate the realized equity premium and eleven financial predictor variables.

Appendix B: Data Description (Continued)

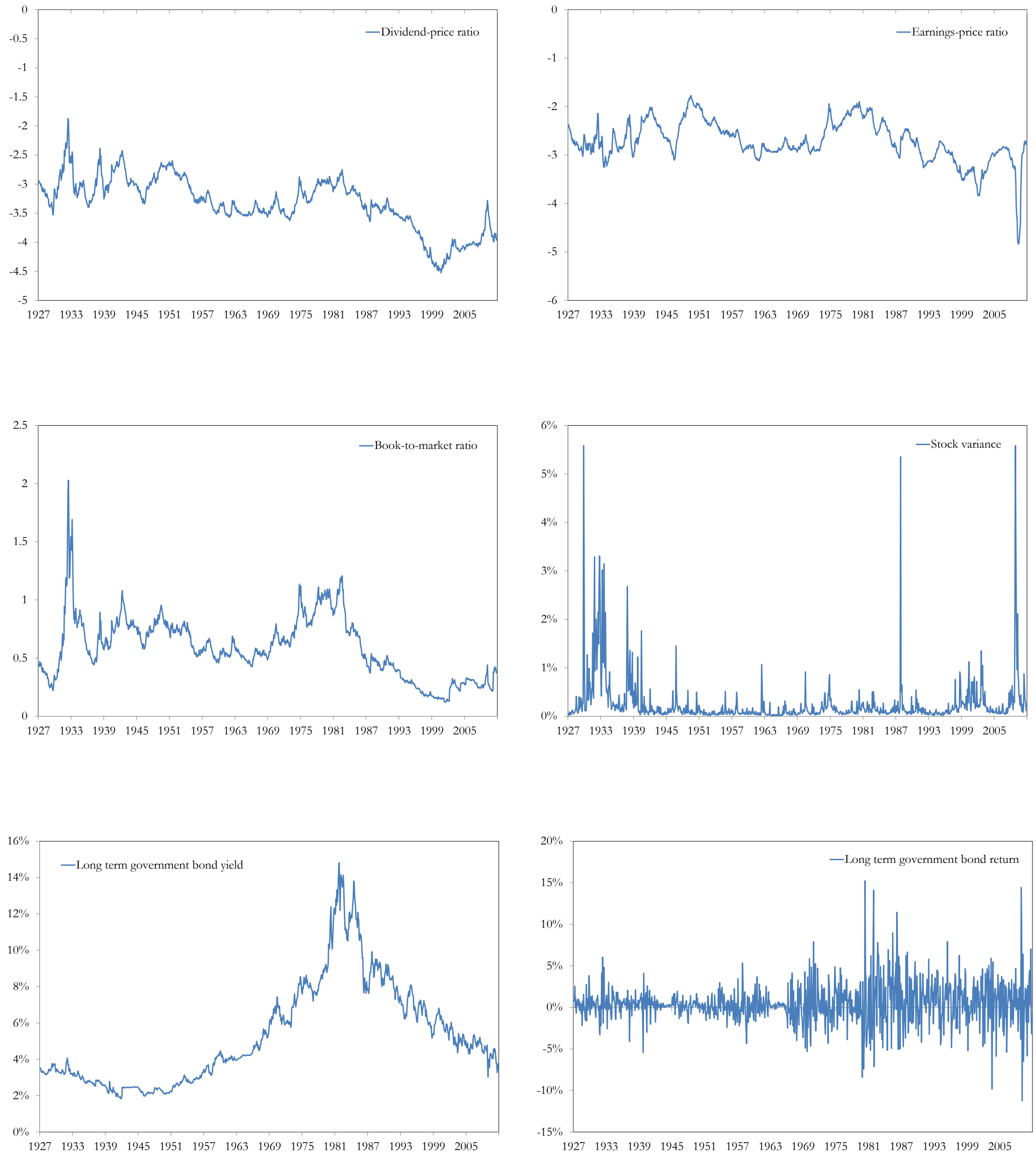


Figure 3.6: *Predictor variables, 1927:1-2010:12.*

Appendix B: Data Description (Continued)

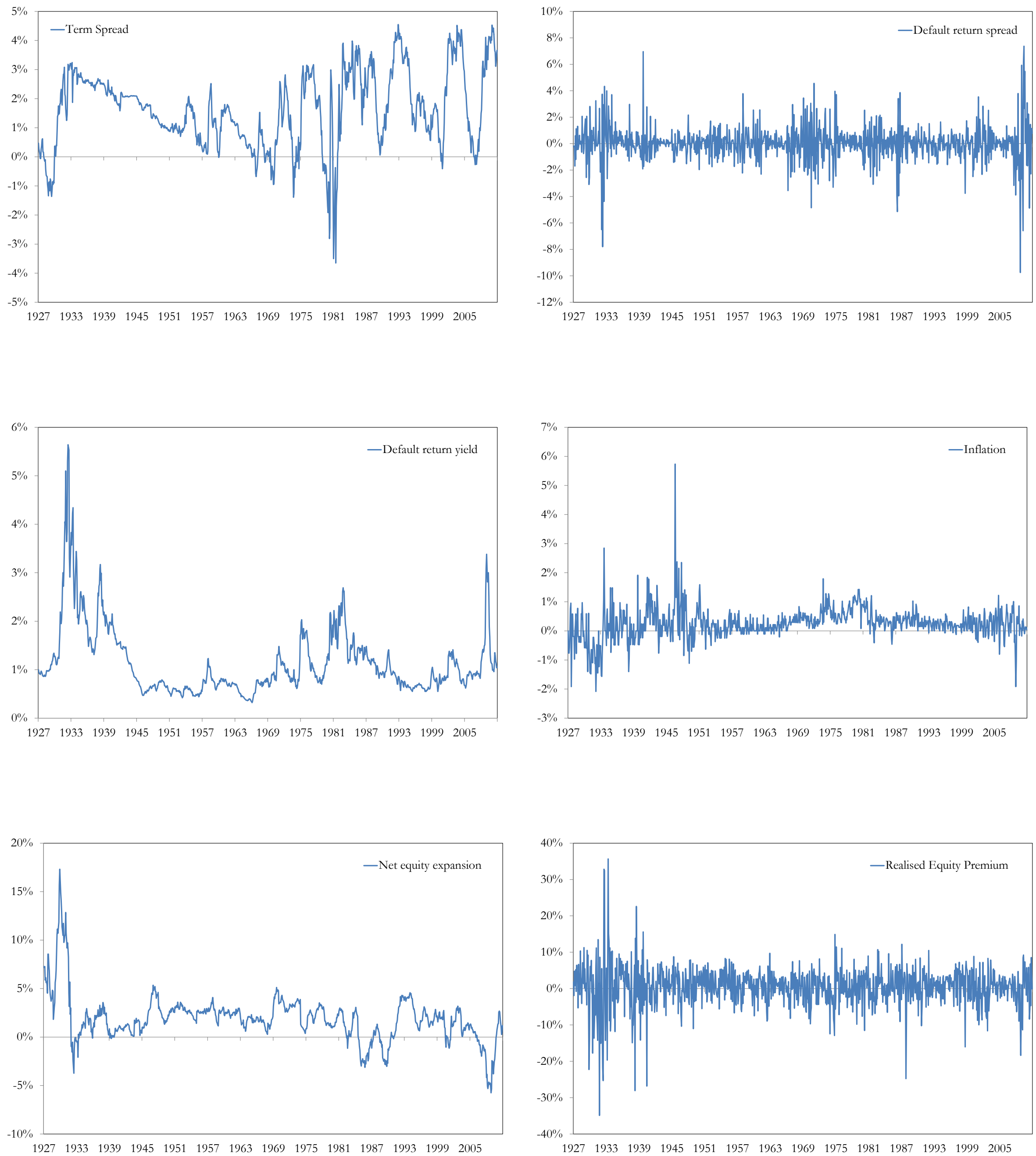


Figure 3.7: *Predictor variables and realized equity return, 1927:1-2010:12.*

Appendix C: In-sample predictability

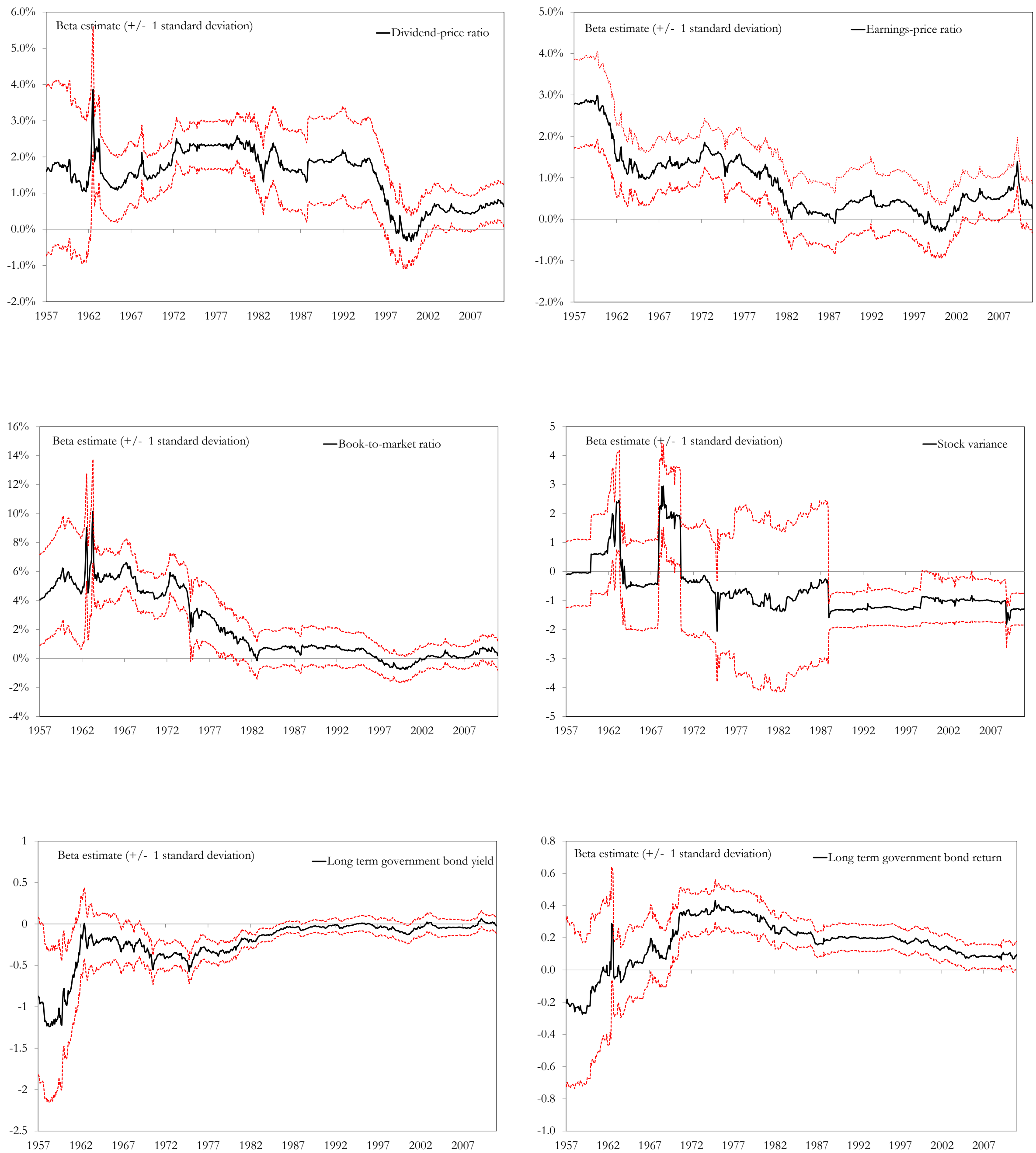


Figure 3.8: *The figures depict rolling beta estimates from single predictor models,* $y_{t+1} = \mathbb{E}[y_{t+1}] + \beta_i(x_{i,t} - \mathbb{E}[x_{i,t}]) + \varepsilon_t$*, based on a rolling windown of 30 years* ($R = 360$)*, 1957:1-2010:12. The dependent variable,* y_{t+1}*, is the CRSP excess return.*

Appendix C: In-sample predictability (Continued)

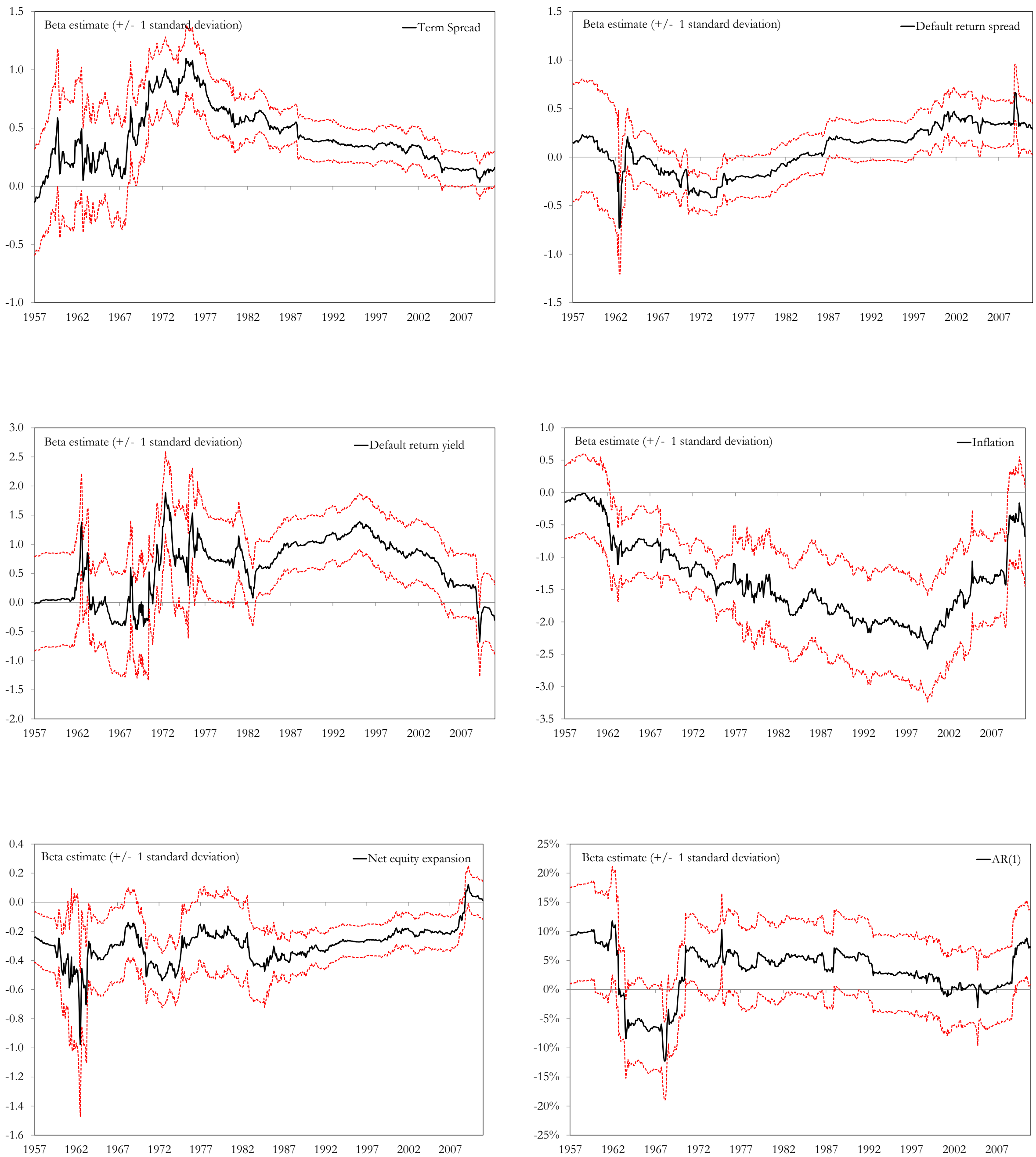


Figure 3.9: *The figures depict rolling beta estimates from single predictor models,* $y_{t+1} = \mathbb{E}[y_{t+1}] + \beta_i(x_{i,t} - \mathbb{E}[x_{i,t}]) + \varepsilon_t$*, based on a rolling windown of 30 years* ($R = 360$)*, 1957:1-2010:12. The dependent variable,* y_{t+1}*, is the CRSP excess return.*

Appendix C: In-sample predictability (Continued)

Figure 3.10: *The figures depict the t-statistic of the rolling beta estimates from a single predictor model,* $y_{t+1} = \mathbb{E}[y_{t+1}] + \beta_i(x_{i,t} - \mathbb{E}[x_{i,t}]) + \varepsilon_t$*, based on a rolling windown of 30 years (*$R = 360$*), 1957:1-2010:12. The dependent variable,* y_{t+1}*, is the CRSP excess return.*

Appendix C: In-sample predictability (Continued)

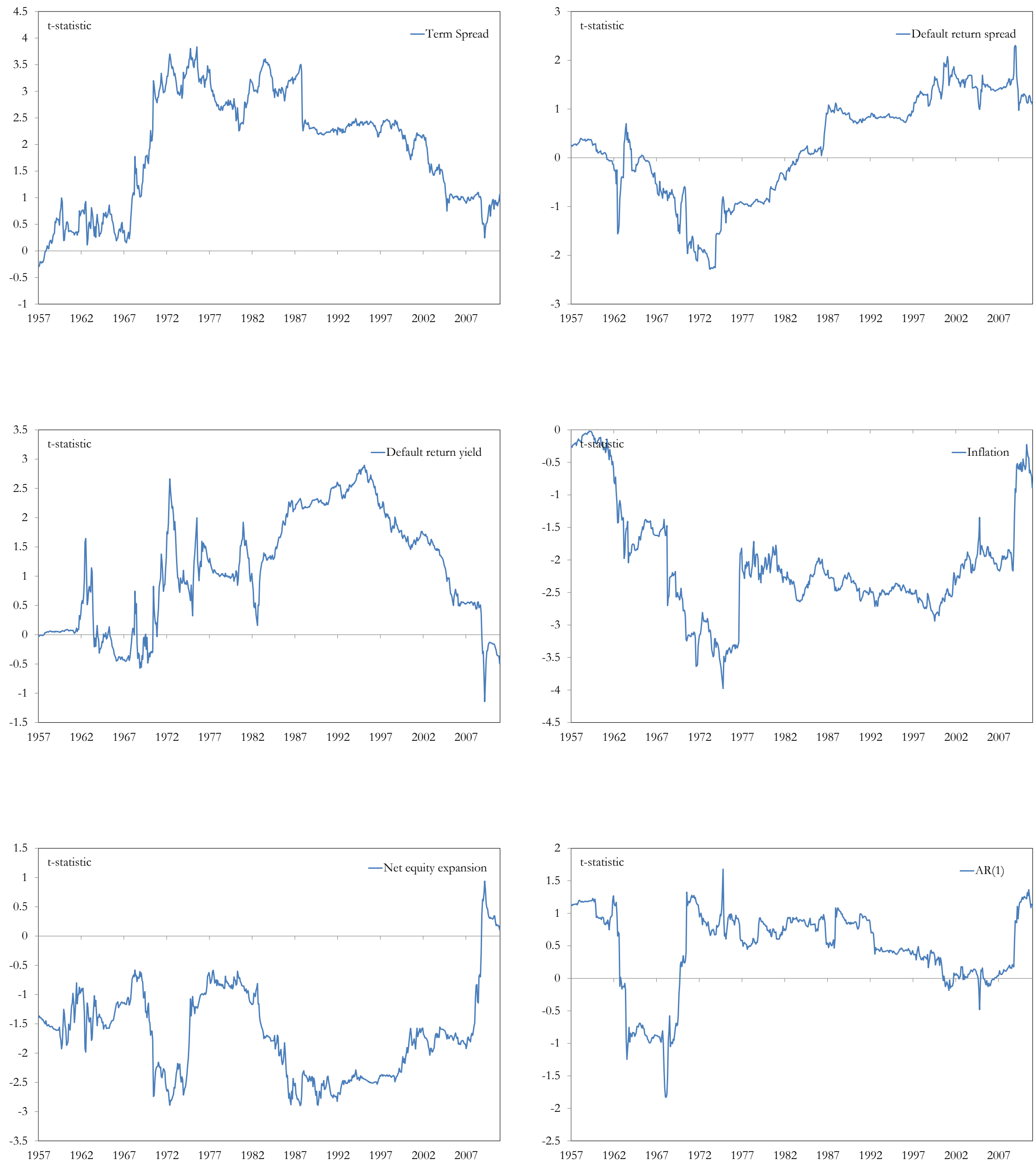


Figure 3.11: *The figures depict the t-statistic of the rolling beta estimates from a single predictor model,* $y_{t+1} = \mathbb{E}[y_{t+1}] + \beta_i(x_{i,t} - \mathbb{E}[x_{i,t}]) + \varepsilon_t$*, based on a rolling windown of 30 years (*$R = 360$*), 1957:1-2010:12. The dependent variable,* y_{t+1}*, is the CRSP excess return.*

Appendix D: Combination Forecasts

A simple portfolio argument motivates the idea of combining forecasts. Suppose we consider the combination of two forecasts that give rise to the forecast errors $\hat{\epsilon}^A_{t+1|t} = y_{t+h} - \hat{y}^A_{t+1|t}$ and $\hat{\epsilon}^B_{t+1|t} = y_{t+1} - \hat{y}^B_{t+h|t}$. Assuming that the forecasts are unbiased we have $\hat{\epsilon}^A_{t+1|t} \sim (0, \sigma^2_A)$ and $\hat{\epsilon}^B_{t+1|t} \sim (0, \sigma^2_B)$ where $\sigma^2_A = Var[\hat{\epsilon}^A_{t+1|t}]$,$\sigma^2_B = Var[\hat{\epsilon}^B_{t+1|t}]$ and $\sigma_{AB} = \rho_{AB}\sigma_A\sigma_B$ is the covariance between the two forecast errors. If we constrain the weights to sum to one they will have the following structure $(\omega, 1-\omega)$. The forecast errors from the combination forecast takes the form $\hat{\epsilon}^c_{t+1|t} = \omega\epsilon^A_{t+1|t} + (1-\omega)\epsilon^B_{t+1|t}$. The forecast error from the combination forecast has zero mean, independent of the weights, and its variance is given by: $\sigma^2_c(\omega) = \omega^2\sigma^2_A + (1-\omega)^2\sigma^2_B + 2\omega(1-\omega)\sigma_{AB}$. Note that this expression is a convex function in ω and we can find the optimal weights that minimise the variance, $\omega*$[6]. The variance of the optimal combination forecast error is given by

$$\sigma^2_c(\omega^*) = \frac{\sigma^2_A\sigma^2_B(1-\rho^2_{AB})}{\sigma^2_A + \sigma^2_B - 2\rho_{AB}\sigma_A\sigma_B}.$$

It is easy to verify that $\sigma^2_c(\omega^*) \leq min(\sigma^2_A, \sigma^2_B)$. The special case where there are no gains from diversification occur only when $\sigma^2_A = 0 \wedge \sigma^2_B = 0$, $\sigma^2_A = \sigma^2_B \vee \rho = 1$ or $\rho = \sigma_A/\sigma_B$.

In the emprical section (3.4), we found that simple combination methods outperformed the, in theory, optimal combination forecasts. Remember from section 3.3 that the optimal forecast was given by: $\omega^* = \frac{\Omega^{-1}\iota}{\iota'\Omega^{-1}\iota}$ where $\Omega = Var[\hat{\boldsymbol{\epsilon}}_{t+1|t}|\mathcal{F}_t]$. Following Timmermann (2006), suppose that all forecast errors have the same variance, σ^2, and correlation, ρ. Then we have that

$$\begin{aligned}\Omega^{-1} &= \frac{1}{\sigma^2(1-\rho)}(\mathbf{I}_N - \frac{\rho}{1+(N-1)\rho}\iota\iota'), \\ & \frac{1}{\sigma^2(1-\rho)(1+(N-1)\rho}((1+(N-1)\rho)\mathbf{I_N} - \rho\iota\iota'),\end{aligned}$$

[6]Differentiation with respect to ω and solving the first order condition we get that: $\omega^* = \frac{\sigma^2_B - \sigma_{AB}}{\sigma^2_A + \sigma^2_B - 2\sigma_{AB}}$ and $(1-\omega^*) = \frac{\sigma^2_A - \sigma_{AB}}{\sigma^2_A + \sigma^2_B - 2\sigma_{AB}}$. Intuitively, a greater weight is assigned to models producing more precise forecasts.

Empirical covariance matrix of the eleven single predictor models' forecast errors, 1957:1-2010:12:

	D/P	E/P	B/M	SVAR	LTY	LTR	TMS	DFR	DFY	INFL	NTIS
D/P	0.188%	0.189%	0.189%	0.186%	0.185%	0.185%	0.185%	0.188%	0.188%	0.185%	0.187%
E/P	0.189%	0.190%	0.190%	0.188%	0.187%	0.186%	0.186%	0.189%	0.189%	0.186%	0.188%
B/M	0.189%	0.190%	0.191%	0.187%	0.186%	0.186%	0.186%	0.188%	0.189%	0.186%	0.187%
SVAR	0.186%	0.188%	0.187%	0.188%	0.186%	0.184%	0.184%	0.187%	0.187%	0.185%	0.186%
LTY	0.185%	0.187%	0.186%	0.186%	0.190%	0.184%	0.185%	0.187%	0.186%	0.187%	0.186%
LTR	0.185%	0.186%	0.186%	0.184%	0.184%	0.187%	0.183%	0.185%	0.186%	0.184%	0.185%
TMS	0.185%	0.186%	0.186%	0.184%	0.185%	0.183%	0.187%	0.185%	0.186%	0.184%	0.184%
DFR	0.188%	0.189%	0.188%	0.187%	0.187%	0.185%	0.185%	0.190%	0.188%	0.186%	0.187%
DFY	0.188%	0.189%	0.189%	0.187%	0.186%	0.186%	0.186%	0.188%	0.190%	0.186%	0.188%
INFL	0.185%	0.186%	0.186%	0.185%	0.187%	0.184%	0.184%	0.186%	0.186%	0.187%	0.185%
NTIS	0.187%	0.188%	0.187%	0.186%	0.186%	0.185%	0.184%	0.187%	0.188%	0.185%	0.188%

Descriptive statistics of the variances:

Average:	0.189%
Median:	0.188%
Stdev:	0.0014%
Min:	0.187%
Max:	0.191%

Descriptive statistics of the correlations:

Average:	98.61%
Median:	98.52%
Stdev:	0.508%
Min:	97.59%
Max:	99.72%

Table 3.3: *Empirical covariance matrix of the eleven single predictor models' forecast errors, 1957:1-2010:12 (P=492)*

where $\mathbf{I}_N$ is a $N \times N$ identity matrix. If we insert this expression into ω^* we have that

$$\begin{aligned} \Omega^{-1}\iota &= \frac{\iota}{\sigma^2(1+(N-1)\rho)} \\ (\iota'\Omega^{-1}\iota)^{-1} &= \frac{\sigma^2(1+(N-1)\rho)}{N} \end{aligned}$$

such that $\omega^* = \left(\frac{1}{N}\right)\iota$. Hence, equal weights are optimal if all forecast errors have the same variance, σ^2, and correlation, ρ. Table 3.3 illustrates the empirical covariance matrix of the eleven single predictor models' forecast errors. This table is helpful in order to understand the poor performance of the optimal combination forecast. Firstly, note that the assumption of equal variance and correlation for all the forecast errors seems to be a good approximation. Based on the previous analytical result we know that this implies that the potential gains, from using optimal weights,

$$\omega^{*'}\Omega\omega^{*} - \left(\frac{1}{N^2}\right)\iota'\Omega\iota,$$

are small. Secondly, note that all the forecasts are highly correlated such that the covariance matrix is almost singular. This implies that the optimal weights will be very sensitive to changes in the covariance matrix. This combined with the fact that we estimate the forecast errors, hereby creating an error-in-variables problem, makes this approach problematic. Even a small estimation error will imply a large change in the optimal combination. In other words, the potential gains from estimating the optimal weights are limited and the potential costs of estimation errors are huge. Therefore, based on this analysis, it is not surprising that the forecast combination based on the *estimated* optimal weights perform poorly.

This analysis also explains why the simple average combination forecast perform well. The assumption of equal variance and correlation seem to be a good approximation, based on table 3.3, and we showed that under these assumptions that equal weights are optimal.

Appendix E: Critical values for the McCracken distribution

k	%-ile	π										
		0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
1	0.99	2.326	1.799	1.604	1.447	1.340	1.221	1.179	1.098	1.021	0.969	0.882
	0.95	1.645	1.117	0.970	0.859	0.722	0.651	0.575	0.510	0.455	0.382	0.334
	0.90	1.280	0.776	0.637	0.530	0.401	0.317	0.246	0.180	0.136	0.116	0.078
2	0.99	2.326	1.757	1.504	1.325	1.180	1.165	0.996	0.953	0.883	0.744	0.640
	0.95	1.645	1.105	0.884	0.753	0.631	0.484	0.401	0.304	0.235	0.166	0.103
	0.90	1.280	0.755	0.569	0.425	0.280	0.155	0.111	0.026	-0.050	-0.094	-0.140

Table 3.4: Percentiles of Non-Standard distribution from McCracken (2007)

This table contains percentiles for the non-standard distribution described in McCracken (2007). The results are based on a rolling estimation scheme. The non-standard distribution depends on the limiting behaviour of the number of forecast relative to the number of observations in the estimation period, $\pi = lim_{P,R\rightarrow\infty}\frac{P}{R}$, and the number of additional variables in the largest model, k.

3.9 Bibliography

Ang, A. & Bekaert, G. (2007), 'Stock return predictability: Is it there?', *Review of Financial Studies* **20**(3), 651–707.

Ball, R. (1978), 'Anomalies in relationships between securities' yields and yield-surrogates', *Journal of Financial Economics* **6**(2), 103–126.

Bates, J. & Granger, C. (1969), 'The combination of forecasts', *Journal of the Operational Research Society* **20**(4), 451–468.

Boudoukh, J., Michaely, R., Richardson, M. & Roberts, M. R. (2007), 'On the importance of measuring payout yield: Implications for empirical asset pricing', *The Journal of Finance* **62**(2), 877–915.

Butler, A. W., Grullon, G. & Weston, J. P. (2005), 'Can managers forecast aggregate market returns?', *The Journal of Finance* **60**(2), 963–986.

Campbell, J. Y. (1987), 'Stock returns and the term structure', *Journal of financial economics* **18**(2), 373–399.

Campbell, J. Y. (2008), 'Viewpoint: Estimating the equity premium', *Canadian Journal of Economics/Revue canadienne d'économique* **41**(1), 1–21.

Campbell, J. Y. & Cochrane, J. H. (1999), 'By force of habit: A consumption-based explanation of aggregate stock market behavior', *The Journal of Political Economy* **107**(2), 205–251.

Campbell, J. Y. & Shiller, R. J. (1988*a*), 'The dividend-price ratio and expectations of future dividends and discount factors', *Review of financial studies* **1**(3), 195–228.

Campbell, J. Y. & Shiller, R. J. (1988*b*), 'Stock prices, earnings, and expected dividends', *The Journal of Finance* **43**(3), 661–676.

Campbell, J. Y. & Thompson, S. B. (2008), 'Predicting excess stock returns out of sample: Can anything beat the historical average?', *Review of Financial Studies* **21**(4), 1509–1531.

Campbell, J. Y. & Yogo, M. (2006), 'Efficient tests of stock return predictability', *Journal of financial economics* **81**(1), 27–60.

Dangl, T. & Halling, M. (2012), 'Predictive regressions with time-varying coefficients', *Journal of Financial Economics* **106**, 157–181.

Dow, C. H. (1920), *Scientific stock speculation*, Magazine of Wall Street.

Ferreira, M. A. & Santa-Clara, P. (2011), 'Forecasting stock market returns: The sum of the parts is more than the whole', *Journal of Financial Economics* **100**(3), 514–537.

Giacomini, R. & White, H. (2006), 'Tests of conditional predictive ability', *Econometrica* **74**(6), 1545–1578.

Goyal, A. & Welch, I. (2003), 'Predicting the equity premium with dividend ratios', *Management Science* **49**(5), 639–654.

Guo, H. (2006), 'On the out-of-sample predictability of stock market returns', *The Journal of Business* **79**(2), 645–670.

Hansen, B. E. (2013), Efficient shrinkage in parametric models. Working Paper, University of Wisconsin Madison.

Hansen, P. R. & Lunde, A. (2006), 'Consistent ranking of volatility models', *Journal of Econometrics* **131**(1), 97–121.

Hodrick, R. J. (1992), 'Dividend yields and expected stock returns: Alternative procedures for inference and measurement', *Review of Financial studies* **5**(3), 357–386.

James, W. & Stein, C. (1961), Estimation with quadratic loss, *in* 'Proceedings of the fourth Berkeley symposium on mathematical statistics and probability', Vol. 1, pp. 361–379.

Keim, D. B. & Stambaugh, R. F. (1986), 'Predicting returns in the stock and bond markets', *Journal of financial Economics* **17**(2), 357–390.

Kothari, S. & Shanken, J. (1997), 'Book-to-market, dividend yield, and expected market returns: A time-series analysis', *Journal of Financial Economics* **44**(2), 169–203.

Lamont, O. (1998), 'Earnings and expected returns', *The Journal of Finance* **53**(5), 1563–1587.

Lehmann, E. L. & Casella, G. (1998), *Theory of point estimation*, Vol. 31, Springer.

Lettau, M. & Van Nieuwerburgh, S. (2008), 'Reconciling the return predictability evidence', *Review of Financial Studies* **21**(4), 1607–1652.

McCracken, M. W. (2007), 'Asymptotics for out of sample tests of granger causality', *Journal of Econometrics* **140**(2), 719–752.

Nelson, C. R. (1976), 'Inflation and rates of return on common stocks', *The Journal of Finance* **31**(2), 471–483.

Pástor, L. & Stambaugh, R. F. (2009), 'Predictive systems: Living with imperfect predictors', *The Journal of Finance* **64**(4), 1583–1628.

Patton, A. J. (2011), 'Volatility forecast comparison using imperfect volatility proxies', *Journal of Econometrics* **160**(1), 246–256.

Rapach, D. E., Strauss, J. K. & Zhou, G. (2010), 'Out-of-sample equity premium prediction: Combination forecasts and links to the real economy', *Review of Financial Studies* **23**(2), 821–862.

Stambaugh, R. F. (1999), 'Predictive regressions', *Journal of Financial Economics* **54**(3), 375–421.

Stein, C. (1956), Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *in* 'Proceedings of the Third Berkeley symposium on mathematical statistics and probability', Vol. 1, pp. 197–206.

Stock, J. H. & Watson, M. W. (2012), 'Generalized shrinkage methods for forecasting using many predictors', *Journal of Business & Economic Statistics* **30**(4), 481–493.

Timmermann, A. (2006), 'Forecast combinations', *Handbook of economic forecasting* **1**, 135–196.

Welch, I. & Goyal, A. (2008), 'A comprehensive look at the empirical performance of equity premium prediction', *Review of Financial Studies* **21**(4), 1455–1508.

West, K. D. (1996), 'Asymptotic inference about predictive ability', *Econometrica: Journal of the Econometric Society* **64**, 1067–1084.

West, K. D., Edison, H. J. & Cho, D. (1993), 'A utility-based comparison of some models of exchange rate volatility', *Journal of International Economics* **35**(1), 23–45.