On the Optimal Performance of Forecasting Systems : The Prequential Approach.

Thesis submitted to the University of London for the Degree of Doctor of Philosophy in the Faculty of Science

by

Konstantinos Skouras

Department of Statistical Science

University College London

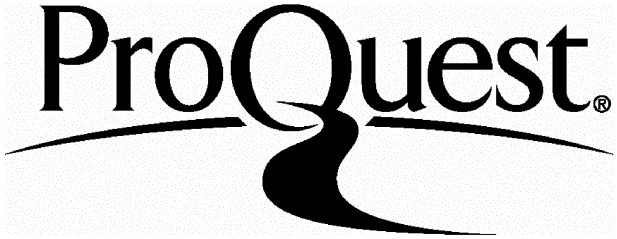
February 1998

ProQuest Number: 10011295

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.



ProQuest 10011295

Published by ProQuest LLC(2016). Copyright of the Dissertation is held by the Author.

All rights reserved.
This work is protected against unauthorized copying under Title 17, United States Code.
Microform Edition © ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

Consider a forecaster who observes a sequence of data on-line and after each new observation makes a forecast (a point estimate or a full probability distribution) for the next observation. A general theory of assessment of such prequential (*pre*dictive-se*quential*) forecasting systems was introduced by Dawid (1984). Within this framework the notion of efficiency of probability forecasting systems was introduced, and it was shown that Bayesian probability forecasting systems are efficient.

In this thesis the concept of prequential efficiency is studied further by presenting some new results. We focus especially on a class of non-Bayesian statistical forecasting systems, the plug-in systems, and we study their efficiency. We show that under suitable conditions the plug-in systems are efficient, but we also show, using counterexamples, that for some models no plug-in system is efficient.

Next, we extend the notion of efficiency to point prediction systems. The efficiency of Bayesian point prediction systems is established, and sufficient conditions are presented for the efficiency of plug-in systems. The results are applied to time series forecasting.

By adopting a predictive point of view, we also study the consistency of extremum estimators for possibly misspecified models. We show, using martingale arguments, that an estimator defined as the minimizer of a statistical criterion measuring predictive performance, converges to the value of the parameter indexing the model that issues the "best" one step ahead predictions for the data at hand. In order to prove our results we establish a martingale version of the uniform law of large numbers.

Acknowledgements

First of all, I would like to thank my supervisor and colleague Prof. A. P. Dawid for his invaluable support and guidance during the last 4 years.

I would also like to thank all members of staff and all research students of the Department of Statistical Science for their friendship and support, and for making our department such an enjoyable place to work and study.

I am indebted to the Greek Scholarships Foundation for partially supporting my studies.

Finally, my deepest gratitude goes to my family and my fiancée Maria for their encouragement and understanding. This thesis is dedicated to them.

Contents

Chapter 1

Introduction

The objective of this thesis is to study the asymptotic behaviour of different forecasting systems, using the prequential (*pre*dictive se*quential*) framework which was introduced by Dawid (1984). More specifically we study the efficiency of forecasting systems for probability forecasting and point prediction systems, and also the consistency of predictive rules for possibly misspecified models.

In Chapter 2 we begin by presenting the prequential framework, which is based on the principle that any statistical model should be assessed by the quality of the forecasts it produces for the specific data at hand. For example, we discuss how a joint distribution for the data can be seen as a method of issuing sequentially probability forecasts for the data, and how this view can be extended to arbitrary statistical models, i.e. collections of probability distributions. By converting a statistical model to a probability forecasting system, we can then use probability assessment techniques to assess its validity, and to compare it with another model. In the same chapter the notion of prequential efficiency is presented as an optimality criterion necessary for the successful replacement of a statistical model by a unique probability forecasting system for inferential purposes.

In Chapter 3, we discuss a specific non-Bayesian method of converting a statistical model to a probability forecasting system, the *plug-in* method, and we study

its prequential efficiency. We study statistical models based on countable and uncountable parameter sets separately, and for the uncountable case we present results using two different approaches using the Kullback-Liebler and the χ^2 distances. We also demonstrate, using counterexamples, that the plug-in method can be inefficient for statistical models where the information from the data grows too fast.

In Chapter 4, we introduce the notion of a point prediction system, as a predictive rule that issues one step ahead point predictions. We show how the notion of prequential efficiency can be extended to point prediction, and we study the efficiency of Bayesian and non-Bayesian methods of making point predictions using this framework. We establish, under weak conditions, the efficiency of Bayesian point prediction systems, and we give sufficient conditions for the efficiency of plug-in point prediction systems. We apply the results to probability forecasting using the Brier score, and stochastic regression models.

In Chapter 5 we study the property of asymptotic consistency. By considering a statistical model as a forecasting system, we show how estimators based on the minimization of a predictive penalty, are consistent, in the sense of converging to the model that gives the "best" predictions. We allow our class of models to be misspecified, and we show how our approach overcomes some problems that other theories of consistency under misspecification face in some non-ergodic cases. A basic tool for our results is a martingale uniform law of large numbers which we prove. We end the chapter with some examples which include a proof of the consistency of least squares estimators in nonlinear stochastic regression models.

In Chapter 6, we summarize the results and discuss future lines of research.

Chapter 2

The Prequential Framework

2.1 Introduction

The theoretical framework that we will use for the study of a forecasting system, is the *prequential framework* (predictive sequential) proposed by Dawid (1984) and explored further by him and his co-workers in a series of papers (Dawid, 1991; Seillier-Moiseiwitsch et al., 1992; Dawid, 1992a; Dawid, 1992b; Dawid, 1997).

The prequential framework is not just a framework for the study and evaluation of forecasting systems. It represents a completely new approach to the traditional problems of statistical inference, based on the principle that is more meaningful to make inferential statements in terms of observable quantities, rather than in terms of unobservable components of a model, such as unknown parameters. Different statistical methods can then be assessed by the validity of their forecasts for some observable quantities, and statistical inferential problems such as model testing, model choice, and robustness can be studied from this perspective.

In this chapter we present the prequential framework in detail, together with some new results.

2.2 Probability Forecasting Systems

In order to introduce the framework, assume that a sequence of random quantities $Y = (Y_1, Y_2, ...)$ will be observed, and, at every step $t \geq 1$, after we observe $y^t = (y_1, \ldots, y_t)$, our task is to issue a "forecast" for the next observation Y_{t+1}. The stochastic quantities (Y_t) can be real numbers or vectors, and it is assumed that they are generated sequentially.

The prequential framework is very general and can incorporate different forms of forecasts such as point prediction, mean-variance prediction, predictive confidence intervals etc. (Dawid, 1992b). In Chapters 2 and 3 we focus on probability forecasting, and therefore the forecasts we consider initially are probability distributions, but later it will become clear that other forms of forecasts can be studied within the same framework. In Chapter 4 we discuss the case of point prediction, where the forecasts are scalars or vectors.

The first notion we introduce is that of a *probability forecasting system (PFS)*, which is a rule that associates with any observed set of data y^t a forecast distribution for the next observation Y_{t+1}. Any PFS determines a unique joint probability distribution for Y (Dawid, 1984), and from any distribution for Y we can construct a PFS. For this reason we will identify a PFS with a probability distribution.

We denote a distribution or, equivalently, a PFS for Y by a bold letter, $\mathbf{F}$ for example. The restriction of $\mathbf{F}$ to the first t observations $(Y_1, Y_2, \ldots, Y_t)$ will be denoted by $\mathbf{F}^t$, and the predictive distribution $\mathbf{F}(Y_{t+1}|y^t)$ by $\mathbf{F}_{t+1}$. Although the predictive distribution $\mathbf{F}_{t+1}$ depends on the observed data y^t, we suppress this from the notation for simplicity. We use small bold letters to denote densities with respect to some underlying measure, usually Lebesgue or counting measure. For example the density of $\mathbf{F}^t$ is $\mathbf{f}^t$.

2.3 Prequential Assessment

Faced with a specific forecasting problem, theoretically any PFS can be used to forecast the uncertain quantities (Y_t). Intuitively we understand that some PFS's will perform well and some badly, depending on the problem and the data at hand. It is important therefore to have a method, or a collection of methods, for assessing the performance of a PFS and comparing it with that of another PFS.

The sequential nature of the problem suggests that any assessment method should judge the PFS within sequence, and not between sequences. By that we mean that we are not interested in assessing the expected performance of a PFS (averaged over all possible realisations), since this will include data which were never observed, and forecasts which were never issued. The prequential point of view suggests that a PFS $\mathbf{F}$ should be assessed by a method which compares the realized forecast distributions $(\mathbf{F}_t)$ with the realized outcomes (y_t) of (Y_t), and does not make use of the full structure of $\mathbf{F}$ as a PFS for Y. This last property was proposed by Dawid (1984) as an inferential principle for the assessment of a PFS, and was termed the *prequential principle*.

Many probability assessment techniques that have been developed, especially in the field of meteorology for the assessment of weather forecasters, respect the prequential principle and can be used to assess the performance of a PFS. These include calibration plots, probability integral transforms, scoring rules etc. The field of probability forecasting assessment is reviewed in Dawid (1986).

2.3.1 Scoring Rules

In this thesis we will focus on the assessment of a PFS using scoring rules. Scoring rules are defined as functions $S(Y_t, \mathbf{F}_t)$ of the outcome Y_t, and the forecast probability distribution $\mathbf{F}_t$. Although they may be seen also as gains to be maximized, we will consider them as penalties that the forecaster should minimize. A scoring

rule is called *proper* if

$$E_{\mathbf{F}_t}\{S(Y_t, \mathbf{F}_t)\} \leq E_{\mathbf{F}_t}\{S(Y_t, \mathbf{P}_t)\}, \tag{2.1}$$

for any distributions $\mathbf{F}_t$ and $\mathbf{P}_t$. It is called *strictly proper* if (2.1) holds with inequality when $\mathbf{F}_t \neq \mathbf{P}_t$. The above property means that the expected penalty for the true distribution will be less than or equal to (proper) or strictly less than (strictly proper) the average penalty for any other distribution.

Many different scoring rules can be used for the assessment of a forecast distribution. One important special case is the *logarithmic scoring rule* defined as:

$$S(Y_t, \mathbf{F}_t) := -\log \mathbf{f}_t(Y_t),$$

where $\mathbf{f}_t(Y_t)$ is the density of $\mathbf{F}_t$ with respect to some fixed underlying measure. This is a proper scoring rule.

Another proper scoring rule is the *Brier score*. If the outcome Y_t is discrete, taking values $(a_1, \ldots, a_m)$, then it is defined as

$$S(Y_t, \mathbf{F}_t) := \sum_{i=1}^{m} \{I(Y_t = a_i) - \mathbf{F}_t(Y = a_i)\}^2,$$

where I is the indicator function, and $\mathbf{F}_t(Y = a_i)$ is the probability that Y_t is equal to a_i under the predictive distribution $\mathbf{F}_t$.

Using scoring rules we might measure the actual perfomance of a PFS $\mathbf{F}$ by its cumulative score

$$S(Y^T, \mathbf{F}) = \sum_{t=1}^{T} S(Y_t, \mathbf{F}_t).$$

An assessment based on this cumulative score satisfies the prequential principle, since it involves only the observed outcomes and the sequence of the forecast distributions $\{\mathbf{F}_t\}$.

2.3.2 Prequential Likelihood

Using the logarithmic score, this assessment of the performance of a PFS $\mathbf{F}$ is essentially its *prequential log-likelihood*, which is defined for data $y^T = (y_1, ..., y_T)$

by

$$L_T(y^T, \mathbf{F}) := \sum_{t=1}^{T} \log \mathbf{f}_t(y_t) = \log \mathbf{f}^T(y^T).$$

Any two PFS's, $\mathbf{F}$ and $\mathbf{Q}$, can be compared, in the light of the data y^T, by the difference of their prequential log-likelihoods:

$$\Delta_T(\mathbf{Q}, \mathbf{F}) := L_T(y^T, \mathbf{Q}) - L_T(y^T, \mathbf{F}).$$

We might prefer the PFS $\mathbf{F}$ if the above difference is negative and the PFS $\mathbf{Q}$ if it is positive. In particular if $\Delta_T(\mathbf{Q}, \mathbf{F}) \to \infty$, as $T \to \infty$, we might consider $\mathbf{F}$ ultimately discredited in favour of $\mathbf{Q}$, and the opposite if $\Delta_T(\mathbf{Q}, \mathbf{F}) \to -\infty$. If the difference stays bounded above and below, we cannot definitively distinguish between the two PFS's, which can then be considered equivalent.

If $Y \sim \mathbf{F}$, the prequential likelihood ratio $\exp(\Delta_T(\mathbf{Q}, \mathbf{F}))$ is a martingale, and so, using standard martingale arguments, we can show that, with $\mathbf{F}$-probability one,

$$\lim_{T-\infty} \Delta_T(\mathbf{Q}, \mathbf{F}) < \infty.$$

This implies that the PFS corresponding to a "true" distribution of the data can not (with probability one) be discredited in favour of any other PFS.

2.4 Statistical Forecasting Systems

In most situations the forecaster does not know the true distribution of the data, but he may be able to specify a suitable class of possible distributions. This class is usually formulated in mathematical terms as a parametric family $\mathcal{P} = \{\mathbf{P}_\theta\}$ of distributions, where θ is an unknown parameter taking values in some set Θ.

Two important inferential issues that arise in this case are *model verification*, i.e. the question of the validity of the parametric family $\mathcal{P}$, and *model selection*, i.e. the comparison of the family $\mathcal{P}$ with another family of distributions, say $\mathcal{Q} = \{\mathbf{Q}_\gamma, \gamma \in \Gamma\}$. The prequential approach to these inferential problems is to replace

a family of distributions for Y by a single PFS. The evaluation of the validity of a family of distributions, or its comparison with another family, reduce then to the simpler problems of assessing the validity of a PFS for Y, or the problem of comparing two candidate PFS's for Y respectively.

In this case the forecaster's aim should be to construct a PFS which will, in a suitable sense to be made precise below, perform at least as well as any other possible PFS, for almost all possible values of θ. Preferably such a PFS should use the assumption of the parametric model, together with any information gathered from the data about the unknown parameter θ, in order to issue its next forecast. Such a PFS will be termed a *statistical forecasting system (SFS).*

The question of how to extract information about θ from the data is one of the main issues in statistical inference, and different methodologies exist, based on different philosophical approaches.

The Bayesian approach accepts that our initial uncertainty for θ should be quantified by a prior distribution for θ, and later, after we observe the data, it should be updated using Bayes's theorem. The Bayesian joint distribution for Y is the mixture

$$\mathbf{B} := \int_{\Theta} \mathbf{P}_{\theta}\, \pi(\theta)\, d\theta,$$

where $\pi(\theta)$ is the prior density. The forecast distribution $\mathbf{B}_{t+1} := \mathbf{B}(Y_{t+1}|y^t)$ is constructed by conditioning on the observed data y^t, and is equivalent to a mixture of the forecast distributions of Y_{t+1} under P_θ, using the posterior distribution of θ:

$$\mathbf{B}_{t+1} = \int_{\Theta} \mathbf{P}_{t+1,\theta}\, \pi_t(\theta)\, d\theta,$$

where $\pi_t(\theta)$ is the posterior distribution of θ based on the prior and the data y^t. We will call such a forecasting system a *Bayesian Forecasting System (BFS).*

Another approach, which avoids the specification of a prior distribution for θ, is the *plug-in* (or *estimative*) approach. A plug-in SFS, say $\mathbf{Q}$, is constructed by calculating at every step t an estimate $\hat{\theta}_t$ of θ based on the current data y^t, and then

using it to specify the predictive distribution for the next observation by replacing the unknown parameter θ with $\hat{\theta}_t$, i.e.

$$\mathbf{Q}_{t+1} := \mathbf{P}_{\hat{\theta}_t}(Y_{t+1}|y^t) := \mathbf{P}_{t+1,\hat{\theta}_t}.$$

There are other ways of constructing forecasting systems, e.g. fiducial SFS's (Dawid, 1984). They are all based on different methods of eliminating the unknown parameter θ in order to generate a forecast distribution. In this thesis we restrict our attention to Bayesian and plug-in SFS's.

As was described above, a SFS can incorporate the assumption of the parametric model plus a learning rule, and different rules will result in different SFS's. A SFS can have an inferential as well as a purely predictive use, and for both purposes it is important to identify a class of "optimal" SFS which compare favourably with any other SFS based on the same parametric model. Then, any assessment based on such SFS will not address the efficacy of the learning process, only the model adequacy. For these reasons the following property of a SFS was introduced by Dawid (1984).

We define a SFS $\mathbf{F}$ to be *efficient* if for *any* other PFS $\mathbf{Q}$, with probability one for all θ in Θ, except perhaps for a subset of measure zero,

$$\limsup_{T-\infty} \Delta_T(\mathbf{Q}, \mathbf{F}) = \limsup_{T-\infty} \{L_T(y^T, \mathbf{Q}) - L_T(y^T, \mathbf{F})\} < \infty. \tag{2.2}$$

If Θ is a subset of $\mathbb{R}^p$ we use Lebesgue measure as the underlying measure. If Θ is countable set we use counting measure, which requires that condition (2.2) should hold for all PFS's $\mathbf{Q}$ and all θ in Θ.

The difference $\Delta(\mathbf{Q}, \mathbf{F})$ measures the relative predictive performance of the two SFS's for the data at hand. According to the above definition a SFS is efficient if it is at least as good as (and possibly better than) any other PFS, and therefore it can not be discredited as a valid model for the data. In that sense, it is the best we can do in modelling the true data generation process. Dawid (1984, 1992a) offers further discussion and justification of the notion of prequential efficiency.

It has been shown that a Bayesian statistical forecasting system based on an almost everywhere positive prior density is efficient. An arbitrary SFS is efficient if and only if it is asymptotically equivalent to a BFS (*cf.* Lemma 2.1). Model selection based on the difference of the prequential log-likelihoods of efficient SFS's for the various models extends the method of log-Bayes factors to non-Bayesian models, and leads to consistent model selection (Dawid, 1992a). It is interesting then to consider the efficiency of plug-in SFS's, since they are proposed as non-Bayesian alternative models for the data, which are free from any prior distributions.

Different authors have discussed the use of plug-in SFS's for model selection. Phillips (1996) discussed the plug-in SFS based on the maximum likelihood estimator and presented an informal argument for its asymptotic equivalence with a BFS. In Phillips and Ploberger (1994) the asymptotic equivalence of a BFS and the maximum likelihood estimator plug-in SFS is established for linear stochastic regression models with Gaussian errors. The use of a plug-in SFS for model selection is also related to Rissanen's predictive minimum description length principle (Rissanen, 1986, 1987, 1989). See also Qian, Gabor, and Gupta (1996) for an application to generalised linear model selection.

Our aim in the next chapter is to present a more general and rigorous study of the plug-in SFS's, and to present sufficient conditions for their prequential efficiency. In the next section we present some results which are necessary for the study of the efficiency of a SFS.

2.5 Absolute Continuity and Efficiency

If C is a subset of Θ then we use the expression $\{\mathbf{P}_\theta, C\}$-as for an event that has probability one under $\mathbf{P}_\theta$, for all θ in C, except perhaps for a set of parameter values of measure zero. If $F(x)$ and $G(x)$ are two distributions for a random variable X, and $f(x)$ and $g(x)$ are their densities with respect to some underlying measure μ,

then by $H(F,G)$ we denote the Hellinger distance:

$$H(F,G) = \Big[\int \{\sqrt{f(x)} - \sqrt{g(x)}\}^2 \mu(dx)\Big]^{1/2},$$

by $K(F,G)$ the Kullback-Leibler distance:

$$K(F,G) = E_f\Big\{\log \frac{f(X)}{g(X)}\Big\},$$

and by $\chi^2(F,G)$ the chi-square distance:

$$\chi^2(F,G) = E_f\Big\{1 - \frac{g(X)}{f(X)}\Big\}^2.$$

Although we refer to the Kullback-Leibler and χ^2 as distances, they are not metrics, as usually $K(F,G) \neq K(G,F)$ and $\chi^2(F,G) \neq \chi^2(G,F)$.

Using the fact that any BFS is efficient, Seillier-Moiseiwitsch, Sweeting, and Dawid (1992) proved the following lemma:

Lemma 2.1 *A SFS* $\mathbf{Q}$ *is prequentially efficient if and only if there exists a BFS* $\mathbf{B}$*, based on an almost everywhere positive prior density, which is absolutely continuous with respect to* $\mathbf{Q}$ *(written* $\mathbf{B} \ll \mathbf{Q}$*).*

By relating efficiency of a SFS to absolute continuity of two distributions, results from probability theory can be used to study the efficiency of a forecasting system.

Theorem 2.1 (Kabanov, Liptser, and Shiryayev (1978)) *If* $\mathbf{R}$ *and* $\mathbf{S}$ *are two distributions for* Y*, and for every* t $\mathbf{R}^t \ll \mathbf{S}^t$*, then a necessary and sufficient condition for* $\mathbf{R} \ll \mathbf{S}$ *is*

$$\sum_{t=1}^{\infty} H^2(\mathbf{R}_t, \mathbf{S}_t) < \infty \qquad \mathbf{R}\text{-}a.s.$$

If we combine the previous two results we get the following necessary and sufficient conditions for the prequential efficiency of a SFS $\mathbf{Q}$:

Lemma 2.2 *A SFS* $\mathbf{W}$ *is efficient if and only if there exists a BFS* $\mathbf{B}$ *such that, for all* t*,* $\mathbf{B}^t \ll \mathbf{W}^t$*, and,* $\{\mathbf{P}_\theta, \Theta\}$*-as,*

$$\sum_{t=1}^{\infty} H^2(\mathbf{B}_t, \mathbf{W}_t) < \infty.$$

We can also prove the following more general result:

Lemma 2.3 *Let* $\mathbf{W}$ *be an efficient SFS, and* $\mathbf{Q}$ *be a SFS such that for every* t, $\mathbf{W}^t \ll \mathbf{Q}^t$. *Then* $\mathbf{Q}$ *is efficient if and only if,* $\{\mathbf{P}_\theta, \Theta\}$*-as,*

$$\sum_{t=1}^{\infty} H^2(\mathbf{Q}_t, \mathbf{W}_t) < \infty.$$

Proof of Lemma 2.3. The SFS $\mathbf{W}$ is efficient, and therefore, from Lemma 2.2, there exists a BFS $\mathbf{B}$ such that $\mathbf{B}^t \ll \mathbf{W}^t \ll \mathbf{Q}^t$ for every t, and $\{\mathbf{P}_\theta, \Theta\}$-as

$$\sum_{t=1}^{\infty} H^2(\mathbf{B}_t, \mathbf{W}_t) < \infty.$$

Using the inequality

$$(a-b)^2 \leq 2\,(a^2 + b^2)$$

it is easy to show that

$$H^2(\mathbf{B}_t, \mathbf{Q}_t) \leq 2\,\{H^2(\mathbf{B}_t, \mathbf{W}_t) + H^2(\mathbf{W}_t, \mathbf{Q}_t)\},$$

and therefore if $\{\mathbf{P}_\theta, \Theta\}$-as

$$\sum_{t=1}^{\infty} H^2(\mathbf{W}_t, \mathbf{Q}_t) < \infty,$$

then $\{\mathbf{P}_\theta, \Theta\}$-as

$$\sum_{t=1}^{\infty} H^2(\mathbf{B}_t, \mathbf{Q}_t) < \infty,$$

and sufficiency is established using Lemma 2.2.

Next we prove necessity. When the SFS $\mathbf{Q}$ is efficient, then there exists a BFS $\mathbf{B}$ such that $\mathbf{B}^t \ll \mathbf{Q}^t$ and $\mathbf{B}^t \ll \mathbf{W}^t$ for every t, and $\{\mathbf{P}_\theta, \Theta\}$-as

$$\sum_{t=1}^{\infty} H^2(\mathbf{B}_t, \mathbf{Q}_t) < \infty$$

and

$$\sum_{t=1}^{\infty} H^2(\mathbf{B}_t, \mathbf{W}_t) < \infty.$$

Again we can use the inequality

$$H^2(\mathbf{W}_t, \mathbf{Q}_t) \leq 2\{H^2(\mathbf{B}_t, \mathbf{W}_t) + H^2(\mathbf{B}_t, \mathbf{Q}_t)\}$$

to show that $\{\mathbf{P}_\theta, \Theta\}$-as

$$\sum_{t=1}^{\infty} H^2(\mathbf{W}_t, \mathbf{Q}_t) < \infty.$$

□

The previous lemma shows that, with probability one for almost all θ, the forecast distributions of any two efficient SFS $\mathbf{Q}$ and $\mathbf{W}$ are asymptotically equivalent for the infinite future $\{\mathbf{P}_\theta, \Theta\}$-as, in the strong sense that,

$$\lim_{T\to\infty} \sum_{t=T}^{\infty} H^2(\mathbf{Q}_t, \mathbf{W}_t) = 0.$$

See also Blackwell and Dubins (1962) for a similar result.

Next we present a sufficient condition for the efficiency of a SFS similar to the condition in Lemma 2.3, but based on the Kullback-Leibler and χ^2 distances. This result is useful in cases where the Hellinger distance is difficult to use.

Lemma 2.4 *An SFS* $\mathbf{Q}$ *is efficient if there is an efficient SFS* $\mathbf{W}$ *such that for every* t, $\mathbf{W}^t \ll \mathbf{Q}^t$, *and* $\{\mathbf{P}_\theta, \Theta\}$*-as*

$$\sum_{t=1}^{\infty} d_t(\mathbf{W}_t, \mathbf{Q}_t) < \infty$$

where $d_t(\mathbf{W}_t, \mathbf{Q}_t)$ *can be any of the distances* $K(\mathbf{W}_t, \mathbf{Q}_t)$, $K(\mathbf{Q}_t, \mathbf{W}_t)$, $\chi^2(\mathbf{W}_t, \mathbf{Q}_t)$ *or* $\chi^2(\mathbf{Q}_t, \mathbf{W}_t)$.

Proof of Lemma 2.4. We show that, for any two distributions F and G, when the Kullback-Leibler and the chi-square distances are finite, they are larger than the squared Hellinger distance. Then the result follows from Lemma 2.3. First we show that this is true for the Kullback-Leibler distance. By the definition of the Hellinger distance,

$$1 - \frac{H^2(F,G)}{2} = \int \sqrt{fg}\, d\mu = E_g(\sqrt{f/g}).$$

Also we can show that

$$-2\log E_g(\sqrt{f/g}) \leq E_g\{\log(g/f)\} = K(G,F).$$

Therefore

$$-\log\Big\{1 - \frac{H^2(F,G)}{2}\Big\} \leq \frac{1}{2}K(G,F),$$

and using the inequality

$$x \leq \log\frac{1}{1-x} \qquad (x<1),$$

we have

$$H^2(F,G) \leq K(G,F).$$

The same argument can be used to show that $H^2(F,G) \leq K(F,G)$.

For the chi-square distance:

$$H^2(F,G) = \int(\sqrt{f}-\sqrt{g})^2 d\mu \leq \int(\sqrt{f}-\sqrt{g})^2\frac{(\sqrt{f}+\sqrt{g})^2}{f}d\mu = \chi^2(F,G),$$

and also by symmetry $H^2(F,G) \leq \chi^2(G,F)$. □

Another sufficient condition for the efficiency of a SFS $\mathbf{Q}$ can be given in terms of the Kullback-Leibler distance between the joint distributions $\mathbf{W}^t$ and $\mathbf{Q}^t$.

Lemma 2.5 *Let* $\mathbf{Q}$ *be a SFS. If there exists an efficient forecasting system* $\mathbf{W}$ *such that*

$$\sup_t K(\mathbf{W}^t, \mathbf{Q}^t) < \infty$$

then $\mathbf{Q}$ *is prequentially efficient.*

Proof of Lemma 2.5. When the condition of the Lemma holds, then for every t the distance $K(\mathbf{W}^t, \mathbf{Q}^t)$ is finite, which implies that $\mathbf{W}^t \ll \mathbf{Q}^t$ for every t. If we define $S_T = \sum_{t=1}^T K(\mathbf{W}_t, \mathbf{Q}_t)$, then S_T is a non-negative sub-martingale under $\mathbf{W}$. According to the submartingale convergence theorem if $\sup_T E_{\mathbf{W}}(S_T)$ is finite then,

with probability one under $\mathbf{W}$, S_T converges to a finite limit. It is straightforward to show that

$$E_{\mathbf{W}}\{K(\mathbf{W}_t, \mathbf{Q}_t)\} = K(\mathbf{W}^t, \mathbf{Q}^t) - K(\mathbf{W}^{t-1}, \mathbf{Q}^{t-1}),$$

and therefore

$$E_{\mathbf{W}}(S_T) = K(\mathbf{W}^T, \mathbf{Q}^T).$$

The result is established using Lemma 2.4, and the fact that if an event holds with probability one under $\mathbf{W}$, then it holds $\{\mathbf{P}_\theta, \Theta\}$-as. □

The next lemma will be useful when working with unbounded parameter sets. It allows us to prove efficiency of a SFS by comparing it to SFS's which are efficient for a subfamily of Θ, e.g. Bayesian forecasting systems based on priors with support on a bounded subset of Θ.

Lemma 2.6 *Let $(C_i, i = 1, 2, ...)$ be a countable family of subsets of Θ, such that $\Theta = \bigcup_{i=1}^{\infty} C_i$. A SFS $\mathbf{Q}$ is efficient if and only if for every i there exists a SFS $\mathbf{R}(i)$ which is efficient for the subfamily $\mathcal{P}_i = \{\mathbf{P}_\theta, \theta \in C_i\}$, and $\mathbf{R}(i) \ll \mathbf{Q}$.*

Proof of Lemma 2.6. We will prove only the efficiency of $\mathbf{Q}$ when for every i $\mathbf{R}(i) \ll \mathbf{Q}$, as the other direction is trivial. Let $\mathbf{R}$ be the SFS defined as

$$\mathbf{R} = \sum_{i=1}^{\infty} 2^{-i}\mathbf{R}(i).$$

For any SFS $\mathbf{W}$, the difference of the prequential log-likelihoods $\Delta_t(\mathbf{W}, \mathbf{R})$ converges to a finite limit $\{\mathbf{P}_\theta, C_i\}$-as for every i. Since Θ is the countable union of C_i, $i \geq 1$, it is easy to see that $\Delta_t(\mathbf{W}, \mathbf{R})$ converges to a finite limit $\{\mathbf{P}_\theta, \Theta\}$-as, and therefore the SFS $\mathbf{R}$ is efficient for the whole family $\mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Theta\}$. Since, for every i, $\mathbf{R}(i) \ll \mathbf{Q}$, then $\mathbf{R} \ll \mathbf{Q}$ and the result is established. □

Corollary 2.1 *If Θ is countable, then a SFS $\mathbf{P}$ is prequentially efficient if and only if, for every θ, $\mathbf{P}_\theta \ll \mathbf{P}$.*

Proof of Corollary 2.1. Without loss of generality assume that $\Theta = \{1, 2, \ldots\}$. If we set $C_i = \{i\}$ and apply Lemma 2.6, we have that a SFS $\mathbf{P}$ is efficient if, for every i, $\mathbf{P}_i \ll \mathbf{P}$. Also if $\mathbf{P}$ is efficient then $\mathbf{P}_i \ll \mathbf{P}$ by the countability of the parameter set. □

Chapter 3

Efficiency and Inefficiency of Plug-in SFS's

3.1 Introduction

In this chapter we present a rigorous study of the efficiency and inefficiency of plug-in SFS's, which, as was discussed in §2.4, are proposed as non-Bayesian alternative models for the data generation process.

In section 3.2 we discuss some advantages and disadvantages of the plug-in approach, and in §3.3 we highlight some problems with the first few observations. In §3.4 we present some general results, and then we study separately the case where the parameter set is countable (section 3.5), and uncountable (section 3.6). For the latter case we present two different approaches one based on the Kullback distance (§3.6.1), and one based on the χ^2 distance (§3.6.2). In §3.7 we show, by means of counterexamples, that plug-in SFS's can be inefficient. We discuss briefly the results in §3.8.

3.2 Plug-in Forecasting systems

As in §2.4, a plug-in SFS is generated by replacing, in the predictive distribution $\mathbf{P}_t(\theta)$, the unknown parameter θ with an estimate based on y^t.

A plug-in SFS may appeal to a non-Bayesian statistician, since it is constructed without the need for the specification of a prior distribution for θ. Another attraction is that in most cases a plug-in SFS is easier to use than a Bayesian SFS, since the analytical form of a Bayesian predictive distribution is usually intractable and numerical methods have to be used to approximate it.

A plug-in SFS also has some disadvantages. Most important, the uncertainty of the estimator of θ is not incorporated in the predictive distribution. Replacing the unknown parameter with an estimate is equivalent to accepting the estimate as the true parameter value. The fact that the estimator is a stochastic quantity, with uncertainty attached to it, is not considered. A value for the estimate may have been calculated from 10 or 10,000 observations, but this is considered irrelevant in the construction of the predictive distribution of a plug-in SFS. This may lead to underestimation of the uncertainty of the future observation Y_{t+1}. As we show later, in some cases this attribute can result in the prequential inefficiency of the plug-in SFS's.

Aitchison (1975) considered the same problem in the context of parametric density estimation, and presented examples where the forecast distribution of a Bayesian SFS is uniformly better (for every θ) than the forecast distribution of a plug-in SFS based on the maximum likelihood estimator. His criterion was the expected Kullback-Leibler distance between the true distribution and the forecast distribution of the SFS. Our investigation is different, because we focus on the asymptotic and within-sequence performance of a SFS, and not on its expected performance. Also we study general models, not only independent identically distributed observations. We will show that for the examples presented by Aitchinson,

although the Bayesian SFS's are slightly better than the plug-in SFS's in terms of the average discrepancy from the true distribution, asymptotically their performances are equivalent.

It is obvious that different estimators result in different SFS's, and performance depends on the estimator sequence used. We will relate the estimative properties of an estimator to the efficiency of the SFS it generates, and apply the results to specific estimators and examples.

3.3 Start-Up

In constructing a plug-in SFS we may face some start-up problems. Perhaps for the first few observations the estimator of θ is not defined, since not enough data are available to calculate it, or, although the estimator may exist, the predictive distribution based on it may not have the same support as the true predictive distribution. This may lead to a non-zero probability that the prequential likelihood will be zero.

Example 3.1 A sequence of independent identically distributed Bernoulli observations (Y_t) is to be observed. The probability $P(Y_t = 1) = \theta$, $0 < \theta < 1$. Let $\mathbf{Q}$ be the plug-in SFS based on the maximum likelihood estimator $\hat{\theta}_T = k/T$, where k is the number of 1's in the first T observations. The predictive distribution $\mathbf{Q}_1$ for X_1 is not defined since we have no data to calculate the MLE.

Even if we ignore the above problem, we can also observe that without any modification the SFS $\mathbf{Q}$ will be inefficient since initially, until we have seen at least one 0 and one 1, the MLE estimator takes the value 0 or 1. This means that the forecast distribution gives probability zero to one of the two possible outcomes. □

In practice a plug-in SFS is used only when a sufficiently large sample is available, and a starting value for the estimator can be calculated. Since our interest is

in asymptotic properties, we avoid these anomalies by assuming that there exists a modification of the initial SFS which avoids these problems, and asymptotically issues identical predictions with the initial SFS. For instance, in Example 3.1, when the estimator $\hat{\theta}_T$ is not defined or takes the values 0 or 1, we can replace it with the estimator $\tilde{\theta}_T = (k+1)/(T+1)$, $k \geq 0$. The two SFS's will eventually issue identical forecasts with probability one for every θ. and therefore by studying the efficiency of the modified SFS we study the efficiency of the original SFS $\mathbf{Q}$ conditioned on the event that the MLE estimator is used only when it is defined, and is not equal to 0 or 1. Of course there may be cases where no such modifications exist, and the plug-in SFS is inefficient.

Example 3.2 Assume that we will observe a sequence of independent identically distributed Uniform $[0, \theta]$ observations. $\theta \in (0, 1)$. The support of the forecast distribution of a plug-in SFS depends on the estimator used to construct it, but with non-zero probability (for non-trivial estimators) it will be smaller than the support of the forecast distribution of any BFS. In that case the plug-in SFS is inefficient since a necessary condition for any SFS to be efficient is that, for every t, the support of its predictive density should include that of a Bayesian SFS. □

In order to have a well defined plug-in SFS, in the following sections we assume that for every $t \geq 0$ the estimator $\hat{\theta}_t$ used to construct the SFS exists, takes values in Θ, and is unique. Also we make the assumption that the support of the density $\mathbf{p}_\theta^t$ of Y^t does not depend on the parameter θ, for every t.

3.4 Projections

We now show that an efficient plug-in SFS exists if and only if, for any BFS $\mathbf{B}$, the plug-in SFS based on the Hellinger projection of the forecast distribution $\mathbf{B}_{t+1}$ into the family of predictive distributions $\mathcal{P}_{t+1} = \{\mathbf{P}_{t+1}(\theta), \theta \in \Theta\}$ is efficient. By $\mathbf{P}_{t+1}(\theta)$ or $\mathbf{P}_{t+1,\theta}$ we denote the predictive distribution $\mathbf{P}(Y_{t+1}|y^t, \theta)$.

Lemma 3.1 *Assume that there exists an efficient plug-in system. For a BFS* $\mathbf{B}$*, let* $\mathbf{W}$ *be the plug-in SFS based on the estimator* $\hat{\theta}_t$ *defined by*

$$\hat{\theta}_t = argmin_{s\in\Theta} H\{\mathbf{B}_{t+1}, \mathbf{P}_{t+1}(s)\}.$$

Then $\mathbf{W}$ *is efficient.*

Proof of Lemma 3.1. Let $\mathbf{Q}$, based on the estimator $(\tilde{\theta}_t)$, be efficient. From Lemma 2.2, for every BFS $\mathbf{B}$, $\{\mathbf{P}_\theta, \Theta\}$-as

$$\sum_{t=1}^{\infty} H^2\Big\{\mathbf{B}_t, \mathbf{P}_t(\tilde{\theta}_{t-1})\Big\} < \infty.$$

The result follows from

$$\sum_{t=1}^{T} H^2\Big\{\mathbf{B}_t, \mathbf{P}_t(\hat{\theta}_{t-1})\Big\} \leq \sum_{t=1}^{T} H^2\Big\{\mathbf{B}_t, \mathbf{P}_t(\tilde{\theta}_{t-1})\Big\},$$

and the fact that, for every t, $\mathbf{B}^t \ll \mathbf{W}^t$. □

Lemma 3.1 shows that if there is a value of θ that minimizes the Hellinger distance between the predictive distribution $\mathbf{P}_{t+1,\theta}$ and the Bayesian predictive distribution $\mathbf{B}_{t+1}$, and there is at least one efficient plug-in SFS, then the plug-in SFS based on these Hellinger projections is efficient as well.

Similar lemmas can be proven for the Kullback-Leibler and χ^2-distance. We present the result for the Kullback-Leibler projection of the forecast distribution of a BFS since in this case the projection is easy to compute and has a special form.

Lemma 3.2 *Assume that there is an efficient plug-in SFS* $\mathbf{Q}$*, based on a sequence of estimators* $\tilde{\theta}_t$*, and there is a BFS* $\mathbf{B}$ *such that,* $\{\mathbf{P}_\theta, \Theta\}$*-as:*

$$\sum_{t=1}^{\infty} K\{\mathbf{B}_t, \mathbf{P}_t(\tilde{\theta}_{t-1})\} < \infty.$$

If $\mathbf{W}$ *is the plug-in SFS generated by the sequence of the estimators* $(\hat{\theta}_t)$ *defined as*

$$\hat{\theta}_t = argmin_{s\in\Theta} \int_\Theta K\{\mathbf{P}_{t+1}(\theta), \mathbf{P}_{t+1}(s)\}\, \pi_t\, d\theta, \tag{3.1}$$

then the SFS $\mathbf{W}$ *is efficient.*

Proof of Lemma 3.2. If e_t is an estimator of θ then the Kullback-Leibler distance $K\{\mathbf{B}_{t+1}, \mathbf{P}_{t+1}(e_t)\}$ can be written as

$$K\{\mathbf{B}_{t+1}, \mathbf{P}_{t+1}(e_t)\} = E_{\theta|y^t}[K\{\mathbf{P}_{t+1}(\theta), \mathbf{P}_{t+1}(e_t)\}] - E_{\theta|y^t}[K\{\mathbf{P}_{t+1}(\theta), \mathbf{B}_{t+1}\}].$$

Therefore the estimator $\tilde{\theta}_t$ which minimises the first term of the right hand side of the above equation minimises the Kullback-Leibler distance between any plug-in forecast distribution and $\mathbf{B}_{t+1}$. Consequently

$$K\{\mathbf{B}_{t+1}, \mathbf{P}_{t+1}(\hat{\theta}_t)\} \leq K_{t+1}\{\mathbf{B}_{t+1}, \mathbf{P}_{t+1}(\tilde{\theta}_t)\},$$

and the result is established by Lemma 2.4. □

Note that the estimator $\hat{\theta}_t$ defined in equation (3.1) is the Bayes estimator when the decision problem is the estimation of the predictive distribution $\mathbf{P}_{t+1}(\theta)$, and the loss function is the Kullback distance. It is also the Kullback projection of the Bayesian predictive distribution $\mathbf{B}_{t+1}$ into the family of distributions $\mathcal{P}_{t+1} = \{\mathbf{P}_{t+1}(\theta), \theta \in \Theta\}$, since it can be shown to minimise $K\{\mathbf{B}_{t+1}, \mathbf{P}_{t+1}(s)\}$ over all parameter values $s \in \Theta$.

3.5 Countable Parameter set

Throughout this section we take Θ to be countable. By Corollary 2.1 a sufficient and necessary condition for a plug-in SFS $\mathbf{Q}$ to be efficient is that, with probability one for every θ,

$$\sum_{t=1}^{\infty} H^2\{\mathbf{P}_t(\theta), \mathbf{Q}_t\} < \infty. \tag{3.2}$$

We call an estimator *consistent* if, with probability one for all θ, it is eventually equal to the true parameter value. We will study how the efficiency of a plug-in SFS is related to the consistency properties of the estimator used.

Suppose that a consistent estimator exists. Then the SFS based on it will be efficient, since with probability one eventually the Hellinger distance between the

predictive distribution of the SFS and the true predictive distribution will be zero, and therefore with probability one, for all θ, condition (3.2) holds.

Now a consistent estimator exists if and only if the distributions $\{P_\theta\}$ are mutually singular. For, first assume that a consistent estimator e_t exists. Then the event "$e_t = \theta$ eventually" has probability one under $\mathbf{P}_\theta$, and zero for any other distribution in the family. Therefore the distributions are mutually singular. If on the other hand the distributions are mutually singular, then it is easy to construct a consistent estimator. First we specify a prior density π_θ on θ, such that $\pi_\theta > 0$ and $\sum_\theta \pi_\theta = 1$. Let $\overline{\theta}_t$ be the posterior mode of θ, i.e. the value that maximises the adjusted likelihood $\pi_\theta \cdot \mathbf{p}^t(y^t|\theta)$ (in cases that there are more than one values of θ that maximise the adjusted likelihood, choose any of them). Then it can be shown, as in §6.4 of Dawid (1992a), that the posterior mode is consistent.

We have shown therefore that consistency is a sufficient condition for the efficiency of a plug-in SFS, and how this is related to the mutual singularity of the distributions in the family. But is consistency a necessary condition for prequential efficiency? The answer is negative, unless we add an extra assumption.

Lemma 3.3 *Suppose that, for every θ, with $\mathbf{P}_\theta$-probability one*

$$\liminf_{t\to\infty}[\inf_{s\neq\theta} H\{\mathbf{P}_t(\theta), \mathbf{P}_t(s)\}] > 0.$$

Let $\hat{\theta}_t$ be a sequence of estimators, and $\mathbf{Q}$ the SFS they generate. Then $\mathbf{Q}$ is prequentially efficient if and only if the estimator $\hat{\theta}_t$ is consistent.

Proof of Lemma 3.3. First suppose that the SFS $\mathbf{Q}$ is efficient. We denote by $H_t(\theta_1, \theta_2)$ the Hellinger distance $H\{\mathbf{P}_t(\theta_1), \mathbf{P}_t(\theta_2)\}$ between the predictive distributions, under $\mathbf{P}_{\theta_1}$ and $\mathbf{P}_{\theta_2}$, for Y_t given the data y^{t-1}. Then for every θ (since $\mathbf{P}_\theta \ll \mathbf{Q}$),

$$\mathbf{P}_\theta\Big\{\sum_{t=1}^{\infty} H_t^2(\theta, \hat{\theta}_{t-1}) < \infty\Big\} = 1,$$

and therefore

$$\mathbf{P}_\theta\Big\{H_t^2(\theta, \hat{\theta}_{t-1}) < \epsilon_\theta \text{ eventually}\Big\} = 1.$$

This implies that

$$\mathbf{P}_\theta\Big\{H_t^2(\theta,\hat{\theta}_{t-1}) = 0 \ \text{ eventually}\Big\} = 1,$$

and thus

$$\mathbf{P}_\theta\Big\{\theta = \hat{\theta}_t \text{ eventually}\Big\} = 1.$$

But, if for every θ, $\mathbf{P}_\theta\{\theta = \hat{\theta}_t \text{ eventually}\Big\} = 1$, then

$$\mathbf{P}_\theta\Big\{\sum_{t=1}^{\infty} H_t^2(\theta,\hat{\theta}_{t-1}) < \infty\Big\} = 1$$

and therefore for all θ, $\mathbf{P}_\theta \ll \mathbf{Q}$. □

A corollary of the above Lemma is that for independent identically distributed observations an efficient plug-in SFS must be based on a consistent estimator.

The above discussion has demonstrated the following result.

Lemma 3.4 *Let the parameter set Θ be countable, and the family of distributions (P_θ) be mutually singular. The SFS $\mathbf{W}$ based on the posterior mode $\overline{\theta}_t$ is efficient.*

If Θ has a finite number of elements, say K, and, for every θ, $\pi_\theta = 1/K$, the posterior mode $\overline{\theta}_t$ is the maximum likelihood estimator, and therefore for a finite set of mutually singular distributions the MLE plug-in SFS is always efficient. When Θ is infinite the MLE estimator does not belong to the class of estimators that maximise the adjusted likelihood. The following example shows that the MLE SFS can be inefficient even when the distributions $\{\mathbf{P}_\theta\}$ are mutually singular.

Example 3.3 We will observe a sequence of random variables which take the values 0 and 1. The true model consists of the countable family of distributions defined as follows:

First number all the finite sequences of 0 and 1's as follows:

1 denotes the sequence 0

2 denotes the sequence 1

3	denotes the sequence	00
4	denotes the sequence	01
5	denotes the sequence	10
6	denotes the sequence	11
7	denotes the sequence	000
⋮	⋮	⋮

and so on. Under $\mathbf{P}_k$, $k = 1, 2, ...$, first we will observe the finite sequence number k, (as defined above), and then a sequence of independent identically distributed Bernoulli observations for which the probability of getting a zero is $1/(k+2)$. Any two models in this family of distributions are singular since there is a value t_0 such that for every $t > t_0$ the Hellinger distance between their forecast distributions for Y_{t+1} is constant and larger than zero.

For every step t, there is only a finite number of models such that their forecast distributions for the next observation Y_{t+1} give non-zero probability to both possible outcomes. In order to have well defined plug-in forecasting systems we define the posterior mode and the MLE estimator to be the parameter values that maximise the adjusted likelihood and the likelihood respectively within the set of these values.

For any sequence of positive prior probabilities (π_k), the BFS, and the plug-in SFS based on the posterior mode, are efficient. But the MLE SFS will be inefficient since, regardless of the data y^t, the MLE estimate is always larger than $2^t - 1$ and, as the number of observations tends to infinity the MLE estimator tends to infinity. Inefficiency follows from Lemma 3.3. □

3.6 Uncountable parameter set

In this Section we assume that the parameter set is uncountable, and an open subset of $\mathbb{R}^p$ (or, more generally, having boundary of Lebesgue measure zero). We

present two different ways of studying the efficiency of a plug-in SFS. The first uses the Kullback-Leibler distance, and the second the χ^2 distance. Both approaches establish the efficiency of a plug-in SFS by finding an upper bound for the distance between the predictive distribution of the plug-in SFS and that of a BFS. When the BFS has support on the whole set Θ, and Θ is unbounded, one usually has to use strong conditions on the tail behaviour of the posterior distributions of θ in order to achieve this bound. Using Lemma 2.6, we see that it is sufficient to find a countable cover $\{C_i\}$ of the parameter set Θ, and for every i, to compare the plug-in SFS with a BFS with support on C_i. This implies that it is sufficient for our purposes to work on an appropriate subset of Θ, which we denote by C, and to give sufficient conditions for the efficiency of the plug-in SFS for the subfamily $\mathbf{P}_C = \{\mathbf{P}_\theta, \theta \in C\}$. When we apply the results, we can choose the cover $\{C_i\}$ in a suitable way in order to show the efficiency of the plug-in SFS for the whole family Θ. Any BFS we use in this section is based on a prior density which has positive support on C, and is equal to zero elsewhere.

3.6.1 Kullback-Leibler Distance

The first approach is based on Lemma 2.5, which shows that a plug-in SFS $\mathbf{Q}$ is efficient if the Kullback-Leibler distance between the joint distributions of a BFS $\mathbf{B}$ and $\mathbf{Q}$ for the first t observations Y^t stays finite as t tends to infinity. The distance $K(\mathbf{B}^t, \mathbf{Q}^t)$ can be decomposed as the difference of two terms :

$$\begin{aligned} K(\mathbf{B}^t, \mathbf{Q}^t) &= \int_C \{K(\mathbf{P}_\theta^t, \mathbf{Q}^t) - K(\mathbf{P}_\theta^t, \mathbf{B}^t)\}\, \pi(\theta)\, d\theta \\ &= E_\pi\{K(\mathbf{P}_\theta^t, \mathbf{Q}^t)\} - E_\pi\{K(\mathbf{P}_\theta^t, \mathbf{B}^t)\}, \end{aligned} \tag{3.3}$$

where the last expectations are with respect to the prior distribution. Since for any forecasting system $\mathbf{W}$

$$E_\pi\{K(\mathbf{P}_\theta^t, \mathbf{W}^t)\} = \sum_{j=1}^{t} E_\pi[E_\theta\{K(\mathbf{P}_{j,\theta}, \mathbf{W}_j)\}],$$

the two terms in equation (3.3) can be interpreted as the overall Bayes risks of the two SFS's $\mathbf{Q}$ and $\mathbf{B}$ for the estimation of the predictive distributions $(\mathbf{P}_{1,\theta}, \ldots, \mathbf{P}_{t,\theta})$ using the Kullback-Leibler distance as the loss function. It easy to show that the BFS $\mathbf{B}$ achieves the minimum Bayes risk, and therefore a SFS is efficient if its Bayes risk is sufficiently close to the minimum risk achieved by $\mathbf{B}$.

The second term $E_\pi\{K(\mathbf{P}_\theta^t, \mathbf{B}^t)\}$ also has many other interpretations. It is the Kullback-Leibler distance between the joint density $\pi(\theta)\,\mathbf{p}_\theta^t(Y^t)$, and the product of marginals $\pi(\theta)$ and $\mathbf{b}^t$. This quantity is the Shannon mutual information between the parameter θ and the sample $Y_1, \ldots Y_t$, and also the expected Kullback-Leibler distance between the posterior and prior densities of θ. In Information Theory it is also the minimal average redundancy of a code (Clarke and Barron, 1994). We denote this quantity by $I(C, Y^t)$, i.e.

$$I(C, Y^t) := \int_C K(\mathbf{P}_\theta^t, \mathbf{B}^t)\,\pi(\theta)\,d\theta,$$

suppressing its dependence on the prior density from the notation.

The mutual information $I(C, Y^t)$ is a quantity that has been well studied for smooth models and independent identically distributed observations (Ibragimov and Hasminskii (1973), Clarke (1989), Clarke and Barron (1990), Clarke and Barron (1994)). Under weak conditions

$$I(C, Y^t) = \frac{p}{2}\log(t) + O(1), \tag{3.4}$$

where p is the dimension of C. As in the decomposition of $K(\mathbf{B}^t, \mathbf{Q}^t)$ the mutual information appears with a negative sign, it will be sufficient to establish that

$$\liminf_{t\to\infty}\left\{I(C, Y^t) - \frac{p}{2}\log(t)\right\} > -\infty. \tag{3.5}$$

Following Clarke (1989), page 76, it can be shown that, for the above result to hold, it is sufficient that there be an estimator e_t such that

$$\limsup_{t\to\infty} \det[E_\pi E_\theta\{t\,(\theta - e_t)(\theta - e_t)'\}] < \infty.$$

On the other hand, the risk $E_\pi E_\theta\{K_{t+1}(\theta,\hat{\theta}_t)\} := E_\pi E_\theta\{K(\mathbf{P}_{t+1,\theta},\mathbf{P}_{t+1,\hat{\theta}_t})\}$ has been studied by Cencov (1981), who showed that, for smooth models and independent identically distributed observations, if $\hat{\theta}_t$ is the maximum likelihood estimator then

$$E_\pi E_\theta\{K_{t+1}(\theta,\hat{\theta}_t)\} < \frac{p}{2t} + O(t^{-3/2}). \tag{3.6}$$

If $\mathbf{Q}$ is a plug-in SFS, based on an arbitrary estimator $\hat{\theta}_t$, which achieves the bound (3.6), then

$$E_\pi E_\theta\{K(\mathbf{P}_\theta^t,\mathbf{Q}^t)\} = E_\pi E_\theta\{\sum_{j=1}^{t} K_j(\theta,\hat{\theta}_{j-1})\} < \frac{p}{2}\log(t) + O(1). \tag{3.7}$$

It follows from equations (3.3), (3.5), and (3.7) that the SFS $\mathbf{Q}$ is efficient since $K(\mathbf{B}^t,\mathbf{Q}^t) = O(1)$.

Using the same arguments we can show the following result which can be applied not only to independent identically distributed observations, but to any parametric family and any estimator for which the assumptions hold.

Theorem 3.1 *Assume that there is a BFS* $\mathbf{B}$ *such that the mutual information* $I(C,Y^t)$ *is lower bounded as in equation (3.5). Let* $\hat{\theta}_t$ *be an estimator which satisfies the following condition:*

$$E_\pi E_\theta\{K_{t+1}(\theta,\hat{\theta}_t)\} < \frac{p}{2t} + b_t$$

where b_t *is a sequence such that* $\sum_{t=1}^{\infty} b_t < \infty$. *If* $\mathbf{Q}$ *is the SFS based on* $\hat{\theta}_t$, *then* $\mathbf{Q}$ *is efficient for the subfamily* $\mathbf{P}_C = (\mathbf{P}_\theta, \theta \in C)$.

Example 3.4 Assume that under $\mathbf{P}_\theta$ the sequence of observations (Y_t) are independent identically distributed having a Normal distribution with mean θ and variance σ^2, known. Let $\Theta = \mathbb{R}$, $C = \Theta$, and $\mathbf{B}$ be the BFS based on a $N(0,1)$ prior. Then

$$I(C,Y^t) = \frac{\log(t+1)}{2},$$

and if the estimator $\hat{\theta}_t$ is the maximum likelihood estimator based on Y^t, then

$$E_\theta\{K_{t+1}(\theta, \hat{\theta}_t)\} = \frac{1}{2t}.$$

It follows that the plug-in SFS based on the maximum likelihood estimator is efficient from Theorem 3.1. This result is in contrast to the result of Aitchison (1975), who showed that some BFS's have a uniformly (for all θ) smaller risk than the MLE plug-in SFS when the loss function is the Kullback-Leibler distance. These differences in the risk are small, usually of order $O(t^{-2})$, so that the difference in the overall risks (when we sum for all t) stays finite. This means that asymptotically the performances of the two forecasting systems are equivalent. □

The Kullback-Leibler distance is not always finite and this may create some problems, especially in cases when $E\{K_{t+1}(\theta, \hat{\theta}_t)\} = \infty$ for every t. Since we focus on a subset C of the set Θ we can avoid this problem by choosing C in such a way that the Kullback-Leibler distances between the predictive distributions of different parameter values in C are always finite. This by itself does not solve the problem since the estimator $\hat{\theta}_t$ takes values in Θ, and not in C. In cases like these, it may be helpful to construct a second estimator $\hat{\theta}_{t,C}$ which takes values in C, being equal to $\hat{\theta}_t$ whenever $\hat{\theta}_t$ is in C, and is defined so that

$$E_\pi E_\theta\{K_{t+1}(\theta, \hat{\theta}_{t,C})\} < \frac{p}{2t} + b_t,$$

where b_t is a sequence such that $\sum_{t=1}^{\infty} b_t < \infty$. Then we can apply theorem (3.1) to show that the plug-in SFS based on $\hat{\theta}_{t,C}$, say $\mathbf{W}$, is efficient for C. If we can then show that for almost all θ in C

$$\mathbf{P}_\theta(\hat{\theta}_t \in C \text{ , eventually}) = 1,$$

then the SFS $\mathbf{Q}$ based on $\hat{\theta}_t$ will be efficient for C, since $\{\mathbf{P}_\theta, C\}$-as

$$\sum_{t=0}^{\infty} H_{t+1}^2(\hat{\theta}_{t,C}, \hat{\theta}_t) < \infty.$$

In the next example we apply this method to the Poisson case.

Example 3.5 Let $(Y_1, Y_2, \ldots)$ be independent identically distributed observations having a Poisson distribution with unknown parameter $\theta > 0$. In this case the Kullback-Leibler distance between two distributions Poisson(θ_1) and Poisson(θ_2) is equal to:

$$K(\theta_1, \theta_2) = \theta_1 \log(\frac{\theta_1}{\theta_2}) + \theta_2 - \theta_1,$$

and for every θ

$$E_\theta\{K(\theta, \hat{\theta}_t)\} = \infty,$$

since there is a non-zero probability that $\hat{\theta}_t$=0.

Let **Q** be the SFS based on the maximum likelihood estimator $\hat{\theta}_t$ of θ, used of course only when $\hat{\theta}_t > 0$. Let $C = [a, b]$ where $0 < a < b$, let **B** be the BFS based on the uniform prior $\pi(\theta) = 1/(b-a)$ on C, and let $\hat{\theta}_{t,C}$ be an estimator defined as follows:

$$\hat{\theta}_{t,C} = \begin{cases} a & \text{if } t = 0 \text{ or } \hat{\theta}_t < a \\ \hat{\theta}_t & \text{if } \hat{\theta}_t \in [a, b] \\ b & \text{if } \hat{\theta}_t > b\,. \end{cases}$$

Since the true value of θ, and the estimator $\hat{\theta}_{t,C}$ take values in C, we can use a Taylor expansion argument to show that for every θ in C

$$K_{t+1}(\theta, \hat{\theta}_{t,C}) = \theta \cdot \{\log(\theta) - \log(\hat{\theta}_{t,C})\} - (\theta - \hat{\theta}_{t,C}) \leq \frac{1}{2\theta}(\theta - \hat{\theta}_{t,C})^2 + \frac{b}{6a^3}|\theta - \hat{\theta}_{t,C}|^3.$$

Since

$$E_\theta(\theta - \hat{\theta}_{t,C})^2 \leq E_\theta(\theta - \hat{\theta}_t)^2 = \theta/t$$

and

$$\sup_{\theta \in C} E_\theta|\theta - \hat{\theta}_{t,C}|^3 = O(t^{-3/2})$$

we have

$$E_\pi E_\theta\{K_{t+1}(\theta, \hat{\theta}_{t,C})\} < \frac{1}{2t} + O(t^{-3/2}).$$

Since $\sup_t E_\pi E_\theta\{t\,(\theta - \hat{\theta}_t)^2\} < \infty$, we also have that

$$\liminf_{t \to \infty}\{I(C, Y^t) - \frac{1}{2}\log(t)\} > -\infty,$$

and therefore the SFS W based on the estimator $\hat{\theta}_{t,C}$ is efficient for C. This implies that the SFS $\mathbf{Q}$ is also efficient for C as the estimator $\hat{\theta}_t$ is (strongly) consistent. Since $\mathbf{Q}$ is efficient for any subset $C = [a, b]$, then $\mathbf{Q}$ is efficient for the whole family $\Theta = \mathbb{R}^+ = \cup_{i=1}^{\infty}[1/i, i]$. □

3.6.2 Chi-square Distance

Next we present an approach based on the χ^2 distance. In order to be able to use Taylor expansions we make the assumption that the subset C is open and convex. $\mathbf{Q}$ is the plug-in SFS based on an estimator $\hat{\theta}_t$, and $\mathbf{B}$ is a BFS based on a prior $\pi(\theta)$ with support on C and zero everywhere else.

In order to establish that the plug-in SFS $\mathbf{Q}$ is efficient, we have to show that, with probability one for almost all θ in C,

$$\sum_{t=1}^{\infty} H^2(\mathbf{B}_t, \mathbf{Q}_t) < \infty,$$

and next we give sufficient conditions for this to hold. By $E_{t-1,\theta}(\cdot)$ we denote the conditional expectation $E_\theta(\cdot|Y^{t-1})$, and by $\lambda \max A$ the maximum eigenvalue of a matrix A.

CONDITIONS

Condition C1.

The conditional density $\mathbf{p}_t(\theta)$ is twice continuously differentiable with derivatives $D_t^{(1)}(\theta)$ and $D_t^{(2)}(\theta)$. Let

$$d_t^{(1)}(\theta) := \{D_t^{(1)}(\theta)\}' D_t^{(1)}(\theta), \tag{3.8}$$

and

$$d_t^{(2)}(\theta) := \sup_{\{h:h'h=1\}} \{\left|h' D_t^{(2)}(\theta)\, h\right|\}. \tag{3.9}$$

Condition C2.

There exist constants $0 < \delta \leq 1$, and $\epsilon > 0$, such that $\{\mathbf{P}_\theta, C\}$-as

$$\sup_{\theta \in C} E_{t-1,\theta}\Big\{\Big|\frac{d_t^{(1)}(\theta)}{\mathbf{p}_t(\theta)}\Big|^2 + \Big|\frac{d_t^{(2)}(\theta)}{\mathbf{p}_t(\theta)}\Big|^{1+\delta}\Big\} = O(1), \tag{3.10}$$

and

$$\int \sup_{\theta_1,\theta_2 \in C} \Big[|\theta_1 - \theta_2|^{-\epsilon} \sup_{h:h'h=1} \Big|h' \{D_t^{(2)}(\theta_1) - D_t^{(2)}(\theta_2)\}\, h\Big|\Big] dY_t = O(1). \tag{3.11}$$

Condition C3.

The estimator $\hat{\theta}_t$ is strongly consistent for almost all θ in C, and, if we denote by $E_{\theta|Y^t}(\cdot)$ the expectation with respect to the posterior density $\pi_t(\theta)$, then

$$E_{\theta|Y^t}\{\sqrt{t}\,(\theta - \hat{\theta}_t)\} = O(\frac{1}{\log(t)}) \tag{3.12}$$

and

$$E_{\theta|Y^t}\{|\sqrt{t}\,(\theta - \hat{\theta}_t)|^{2+\epsilon}\} = O(1), \tag{3.13}$$

where ϵ is the constant used in condition C2. The above orders should hold $\{\mathbf{P}_\theta, C\}$-as, or in expectation for almost all θ in C, or in expectation under the BFS $\mathbf{B}$.

Theorem 3.2 *Assume that the conditions C1-C3 hold. Then* $\{\mathbf{P}_\theta, C\}$*-as*

$$\sum_{t=1}^{\infty} H^2(\mathbf{B}_t, \mathbf{Q}_t) < \infty,$$

and the SFS $\mathbf{Q}$ *is efficient.*

Comments:

(1) Condition C1 is a smoothness condition. Since we are using this condition in order to bound the Hellinger distances, it is sufficient that the conditional densities $\mathbf{p}_t(\theta)$ are eventually differentiable $\{\mathbf{P}_\theta, C\}$-as.

(2) Condition C2 asks for uniformly bounded derivatives on C. Equation (3.11) describes a condition which is weaker than a uniform bound on the third derivative. We can avoid this condition if we make stronger the condition described in

equation (3.10). Condition C2 is restrictive, in the sense that it does not allow the information from each observation to grow to infinity, but we do not think that it can be relaxed except in special cases. As we will show in the next section, when the information grows very fast the plug-in SFS's are inefficient.

(3) Condition C3 describes the conditions that the estimator $\hat{\theta}_t$ should satisfy. According to equation (3.12) the squared posterior bias of the estimator $\hat{\theta}_t$ should go to zero, at least with rate $O\{1/\log^2(t)\}$, which means that $\hat{\theta}_t$ should be close to the posterior mean of θ (under the Bayesian measure $\mathbf{B}$). Equation (3.13) controls the behaviour of higher moments. There exist results in the bibliography which can be used to verify condition C3 (Johnson, 1970; Crowder, 1988; Ibragimov and Hasminskii, 1980), especially for maximum likelihood estimators.

(4) A different interpretation can be given to condition (3.12) when we consider it in expectation under the BFS $\mathbf{B}$. If by $\tilde{\theta}_t$ we denote the posterior mean of θ under $\mathbf{B}$, then

$$\left[E_{\theta|Y^t}\{\sqrt{t}\,(\theta-\hat{\theta}_t)\}\right]^2 = t\,|\hat{\theta}_t-\tilde{\theta}_t|^2 = E_{\theta|Y^t}\{t\,(\theta-\hat{\theta}_t)^2\} - E_{\theta|Y^t}\{t\,(\theta-\tilde{\theta}_t)^2\}.$$

If we calculate the expectation of the above under $\mathbf{B}$ we have

$$E_\pi E_\theta\left[E_{\theta|Y^t}\{\sqrt{t}\,(\theta-\hat{\theta}_t)\}\right]^2 = \int_C E_\theta(t\,|\theta-\hat{\theta}_t|^2)\,\pi(\theta)\,d\theta - \int_C E_\theta(t\,|\theta-\tilde{\theta}_t|^2)\,\pi(\theta)\,d\theta. \tag{3.14}$$

For an estimator e_t, let

$$R_2(e_t) := E_\pi\{E_\theta(|e_t-\theta|^2)\}.$$

Then equation (3.14) shows that an estimator $\hat{\theta}_t$ can produce an efficient SFS if its risk $R_2(\hat{\theta}_t)$ is sufficiently close to the minimum risk $R_2(\tilde{\theta}_t)$ achieved by the posterior mean $\tilde{\theta}_t$. If we consider $R_2(e_t)$ as a measure of the (Bayesian) estimative efficiency of the estimator e_t, then this verifies Dawid's conjecture (Dawid, 1984) that efficient estimators produce efficient SFS's under suitable regularity conditions. The conjecture does not hold for the classical notion of efficiency which is based on

the variance of the asymptotic distribution of the estimator. See the next section for further discussion.

Before we give the proof of Theorem 3.2 we need the following Lemma,

Lemma 3.5 *Let $f(x)$ be a probability density which is positive on the set S. If $g(x)$ is a non-negative function on S such that*

$$\int_S \left|\sqrt{f(x)} - \sqrt{g(x)}\right|^2 dx < \infty,$$

then for every $0 \leq \epsilon \leq 1$

$$\int_S \left|\sqrt{f(x)} - \sqrt{g(x)}\right|^2 dx \leq E_f \left|\frac{f(x) - g(x)}{f(x)}\right|^{1+\epsilon}.$$

Proof.

$$\int_S \left|\sqrt{f(x)} - \sqrt{g(x)}\right|^2 dx = \int_S \left|\sqrt{f(x)} - \sqrt{g(x)}\right|^{2\epsilon} \left|\sqrt{f(x)} - \sqrt{g(x)}\right|^{2-2\epsilon} dx$$

$$\leq \int_S \frac{\left|\sqrt{f(x)} - \sqrt{g(x)}\right|^{2\epsilon} \left|\sqrt{f(x)} + \sqrt{g(x)}\right|^{2\epsilon}}{\{f(x)\}^{\epsilon}} \left\{\left|\sqrt{f(x)} - \sqrt{g(x)}\right|^2\right\}^{1-\epsilon} dx$$

$$\leq \int_S \frac{\left|f(x) - g(x)\right|^{2\epsilon}}{\{f(x)\}^{\epsilon}} \left|f(x) - g(x)\right|^{1-\epsilon} dx = \int_S \frac{\left|f(x) - g(x)\right|^{1+\epsilon}}{\{f(x)\}^{\epsilon}} dx = E_f \left|\frac{f(x) - g(x)}{f(x)}\right|^{1+\epsilon}. \quad \square$$

Proof of Theorem 3.2. Let $\tilde{\theta}_t$ be the posterior mean of θ:

$$\tilde{\theta}_t = \int_C \theta \, \pi_t(\theta) \, d\theta,$$

and $\mathbf{R}$ the plug-in SFS based on it. By $\mathbf{b}_{t+1}$, $\mathbf{r}_{t+1}$ and $\mathbf{q}_{t+1}$ we denote the predictive densities of the distributions $\mathbf{B}_{t+1}, \mathbf{R}_{t+1}$ and $\mathbf{Q}_{t+1}$ respectively. In order to show that $\{\mathbf{P}_\theta, C\}$-as

$$\sum_{t=1}^{\infty} H^2(\mathbf{B}_t, \mathbf{Q}_t) < \infty,$$

it is sufficient to show that $\{\mathbf{P}_\theta, C\}$-as

$$\sum_{t=1}^{\infty} H^2(\mathbf{B}_t, \mathbf{R}_t) < \infty \qquad \text{and} \qquad \sum_{t=1}^{\infty} H^2(\mathbf{R}_t, \mathbf{Q}_t) < \infty,$$

since $H^2(\mathbf{B}_t, \mathbf{Q}_t) \leq 2\{H^2(\mathbf{B}_t, \mathbf{R}_t) + H^2(\mathbf{R}_t, \mathbf{Q}_t)\}$. We begin with the Hellinger distance $H(\mathbf{B}_{t+1}, \mathbf{R}_{t+1})$. Using the conditions C1 and C2, for every t and every θ in C, we have

$$\begin{aligned}\mathbf{p}_{t+1}(\theta) - \mathbf{p}_{t+1}(\tilde{\theta}_t) &= \{D^{(1)}_{t+1}(\tilde{\theta}_t)\}'(\theta - \tilde{\theta}_t) + (1/2)(\theta - \tilde{\theta}_t)' D^{(2)}(\tilde{\theta}_t)(\theta - \tilde{\theta}_t) \\ &\quad + (1/2)(\theta - \tilde{\theta}_t)' \{D^{(2)}_{t+1}(\theta^*) - D^{(2)}_{t+1}(\tilde{\theta}_t)\}(\theta - \tilde{\theta}_t),\end{aligned}$$

where θ^* is a point which lies on the line joining θ and $\tilde{\theta}_t$. Using the fact that $|\theta - \theta^*| \leq |\theta - \tilde{\theta}_t|$ we have

$$\begin{aligned}&\left|\mathbf{p}_{t+1}(\theta) - \mathbf{p}_{t+1}(\tilde{\theta}_t) - \{D^{(1)}_{t+1}(\tilde{\theta}_t)\}'(\theta - \tilde{\theta}_t) - (1/2)(\theta - \tilde{\theta}_t)' D^{(2)}(\tilde{\theta}_t)(\theta - \tilde{\theta}_t)\right| \\ &\leq (1/2)\,|\theta - \tilde{\theta}_t|^{2+\epsilon}\,\Gamma_{t+1}, \qquad (3.15)\end{aligned}$$

where $\Gamma_{t+1} := \sup_{\theta_1, \theta_2 \in C}\left[|\theta_1 - \theta_2|^{-\epsilon} \sup_{\{h: h'h=1\}}\left|h'\{D^{(2)}_{t+1}(\theta_1) - D^{(2)}_{t+1}(\theta_2)\}h\right|\right]$. Define

$$u_{t+1} := \max\left\{0, \mathbf{p}_{t+1}(\tilde{\theta}_t) + \int_C (1/2)(\theta - \tilde{\theta}_t)' D^{(2)}(\tilde{\theta}_t)(\theta - \tilde{\theta}_t)\,\pi_t(\theta)\,d\theta\right\}.$$

Then

$$\begin{aligned}H^2(\mathbf{B}_{t+1}, \mathbf{R}_{t+1}) &= \int\left\{\sqrt{\mathbf{b}_{t+1}} - \sqrt{\mathbf{p}_{t+1}(\tilde{\theta}_t)}\right\}^2 dy_{t+1} \\ &= \int\left\{\sqrt{\mathbf{b}_{t+1}} - \sqrt{u_{t+1}} + \sqrt{u_{t+1}} - \sqrt{\mathbf{p}_{t+1}(\tilde{\theta}_t)}\right\}^2 dy_{t+1} \\ &\leq 2\int(\sqrt{\mathbf{b}_{t+1}} - \sqrt{u_{t+1}})^2\,dy_{t+1} + 2\int\left\{\sqrt{u_{t+1}} - \sqrt{\mathbf{p}_{t+1}(\tilde{\theta}_t)}\right\}^2 dy_{t+1} \qquad (3.16)\end{aligned}$$

Using Lemma 3.5, equation (3.15), and the fact that

$$\int_C \left\{D^{(1)}_{t+1}(\tilde{\theta}_t)\right\}'(\theta - \tilde{\theta}_t)\Big\}\,\pi_t(\theta)\,d\theta = 0,$$

we have

$$\begin{aligned}&\int(\sqrt{\mathbf{b}_{t+1}} - \sqrt{u_{t+1}})^2\,dy_{t+1} \leq \int|\mathbf{b}_{t+1} - u_{t+1}|\,dy_{t+1} \\ &\leq (1/2)\,\frac{1}{t^{1+\epsilon/2}}\int_C |\sqrt{t}\,(\theta - \tilde{\theta}_t)|^{2+\epsilon}\,\pi_t(\theta)\,d\theta \int \Gamma_{t+1}\,dy_{t+1} \qquad (3.17)\end{aligned}$$

and also

$$\int \left(\sqrt{u_{t+1}} - \sqrt{\mathrm{p}_{t+1}(\tilde{\theta}_t)}\right)^2 dy_{t+1} \leq E_{t,\tilde{\theta}_t} |\frac{u_{t+1} - \mathrm{p}_{t+1}(\tilde{\theta}_t)}{\mathrm{p}_{t+1}(\tilde{\theta}_t)}|^{1+\delta}$$

$$\leq \frac{1}{t^{1+\delta}} \Big\{ \int_C |\sqrt{t}\,(\theta - \tilde{\theta}_t)|^2 \,\pi_t(\theta)\, d\theta \Big\}^{1+\delta} \sup_{\theta \in C} E_{t,\theta} \Big\{ \frac{d^{(2)}_{t+1}(\theta)}{\mathrm{p}_{t+1}(\theta)} \Big\}^{1+\delta} \qquad (3.18)$$

We would like now to show that, $\{\mathbf{P}_\theta, C\}$-as,

$$\sum_{t=1}^{\infty} \int (\sqrt{\mathrm{b}_{t+1}} - \sqrt{u_{t+1}})^2 \, dy_{t+1} < \infty,$$

and also

$$\sum_{t=1}^{\infty} \int \left\{\sqrt{u_{t+1}} - \sqrt{\mathrm{p}_{t+1}(\tilde{\theta}_t)}\right\}^2 dy_{t+1} < \infty,$$

because then we can use equation (3.16) to show that $\sum_{t=1}^{\infty} H^2(\mathbf{B}_t, \mathbf{R}_t) < \infty$. There are two ways to do this, depending on the version of condition C3 we use. If for example we know that $E_{\theta|Y^t}\{|\sqrt{t}\,(\theta - \hat{\theta}_t)|^{2+\epsilon}\}$ is of order $O(1)$ $\{\mathbf{P}_\theta, C\}$-as, then, from condition C2, the terms in equations (3.17) and (3.18) are of order $O(t^{-1-\epsilon/2})$ and $O(t^{-1-\delta})$ $\{\mathbf{P}_\theta, C\}$-as respectively, and then using equation 3.16 there exists a sufficiently small $d > 0$ such that

$$H^2(\mathbf{B}_t, \mathbf{R}_t) = O(\frac{1}{t^{1+d}}) \qquad \{\mathbf{P}_\theta, C\}\text{-as},$$

and therefore, $\{\mathbf{P}_\theta, C\}$-as,

$$\sum_{t=1}^{\infty} H^2(\mathbf{B}_t, \mathbf{R}_t) < \infty. \qquad (3.19)$$

In order to use condition C3 with the orders holding in expectation for almost all θ, or under $\mathbf{B}$, observe that the terms in equations 3.17 and 3.18 are positive, and therefore when you sum them you get a submartingale. In order to show that the sum stays finite $\{\mathbf{P}_\theta, C\}$-as it is sufficient to show that it stays finite in expectation. The same result 3.19 follows on observing that the last factor in each equation is $O(1)$ $\{\mathbf{P}_\theta, C\}$-as from condition C2, and also that if an event has probability one under $\mathbf{B}$, it has probability one for almost all θ in C.

Next, we turn our attention to the Hellinger distance $H(\mathbf{R}_{t+1}, \mathbf{Q}_{t+1})$ which for simplicity we also denote by $H_{t+1}(\tilde{\theta}_t, \hat{\theta}_t)$. We want to show that $\sum_{t=1}^{\infty} H^2(\mathbf{R}_{t+1}, \mathbf{Q}_{t+1}) < \infty$ in order to complete the proof.

From condition C3 we have that $\{\mathbf{P}_\theta, C\}$-as there is a t_0, which may depend on the sequence, such that for all $t > t_0$ the estimator $\hat{\theta}_t$ is in the set C. Then we can use a Taylor expansion. For every $t \geq t_0$ let

$$g_{t+1} = \max[0, \mathrm{p}_{t+1}(\hat{\theta}_t) + \{D^{(1)}_{t+1}(\hat{\theta}_t)\}'(\tilde{\theta}_t - \hat{\theta}_t)].$$

Then using Lemma 3.5 we have

$$\begin{aligned}
H^2_{t+1}(\hat{\theta}_t, \tilde{\theta}_t) &= \int \left\{\sqrt{\mathrm{p}_{t+1}(\hat{\theta}_t)} - \sqrt{\mathrm{p}_{t+1}(\tilde{\theta}_t)}\right\}^2 dy_{t+1} \\
&= \int \left\{\sqrt{\mathrm{p}_{t+1}(\hat{\theta}_t)} - \sqrt{g_{t+1}} + \sqrt{g_{t+1}} - \sqrt{\mathrm{p}_{t+1}(\tilde{\theta}_t)}\right\}^2 dy_{t+1} \\
&\leq 2\int \left\{\sqrt{\mathrm{p}_{t+1}(\hat{\theta}_t)} - \sqrt{g_{t+1}}\right\}^2 dy_{t+1} + 2\int \left\{\sqrt{g_{t+1}} - \sqrt{\mathrm{p}_{t+1}(\tilde{\theta}_t)}\right\}^2 dy_{t+1} \\
&\leq 2\int |\mathrm{p}_{t+1}(\hat{\theta}_t) - g_{t+1}|\, dy_{t+1} + 2\int \frac{\{\mathrm{p}_{t+1}(\tilde{\theta}_t) - g_{t+1}\}^2}{\mathrm{p}_{t+1}(\tilde{\theta}_t)}\, dy_{t+1} \\
&\leq |\hat{\theta}_t - \tilde{\theta}_t|^2 \sup_{\theta \in C} E_{t,\theta}\left|\frac{d^{(2)}_{t+1}(\theta)}{\mathrm{p}_{t+1}(\theta)}\right| + 2\,|\hat{\theta}_t - \tilde{\theta}_t|^2 \sup_{\theta \in C} E_{t,\theta}\left|\frac{d^{(1)}_{t+1}(\theta)}{\mathrm{p}_{t+1}(\theta)}\right|^2 \\
&\leq |\hat{\theta}_t - \tilde{\theta}_t|^2 \left\{\sup_{\theta \in C} E_{t,\theta}\left|\frac{d^{(2)}_{t+1}(\theta)}{\mathrm{p}_{t+1}(\theta)}\right| + 2 \sup_{\theta \in C} E_{t,\theta}\left|\frac{d^{(1)}_{t+1}(\theta)}{\mathrm{p}_{t+1}(\theta)}\right|^2\right\} \\
&\leq \frac{1}{t\,(\log t)^2}\,|\sqrt{t}\,\log(t)\,(\hat{\theta}_t - \tilde{\theta}_t)|^2\, 2 \sup_{\theta \in C} E_{t,\theta}\left\{\left|\frac{d^{(2)}_{t+1}(\theta)}{\mathrm{p}_{t+1}(\theta)}\right| + \left|\frac{d^{(1)}_{t+1}(\theta)}{\mathrm{p}_{t+1}(\theta)}\right|^2\right\} \\
&= O\left\{\frac{1}{t\,(\log t)^2}\right\} \qquad \{\mathbf{P}_\theta, C\}\text{-as}
\end{aligned}$$

using conditions C2 and C3 (where the orders hold $\{\mathbf{P}_\theta, C\}$-as). It follows that $\{\mathbf{P}_\theta, C\}$-as

$$\sum_{t=0}^{\infty} H^2_{t+1}(\hat{\theta}_t, \tilde{\theta}_t) < \infty. \tag{3.20}$$

Again observe that the terms in the above inequalities are all positive, and therefore their sum is a submartingale which implies that condition C3 can be used in expectation.

The theorem has been established since equations (3.19) and (3.20) imply that

$$\sum_{t=1}^{\infty} H^2(\mathbf{B}_t, \mathbf{Q}_t) < \infty.$$

□

Next we present an example in order to show how the theorem (3.2) can be applied in specific cases.

Example 3.6 Let (Y_t) be independent identically distributed observations having an exponential distribution with mean $1/\theta$. Then

$$\mathrm{p}(y_t, \theta) = \theta\, e^{-\theta\, y_t}$$

and

$$\begin{aligned} D_t^{(1)}(\theta) &= \epsilon^{-\theta\, y_t}\,(1 - \theta\, y_t), \\ D_t^{(2)}(\theta) &= y_t\, \epsilon^{-\theta\, y_t}\,(\theta\, y_t - 2) \\ \frac{d^3 p_t(\theta)}{d\theta^3} &= y_t^2\, \epsilon^{-\theta\, y_t}\,(3 - \theta\, y_t). \end{aligned}$$

Then if $C = (a, b)$,

$$\begin{aligned} \frac{d_t^{(1)}(\theta)}{\mathrm{p}_t(\theta)} &= \frac{1}{\theta} - y_t \\ \frac{d_t^{(2)}(\theta)}{\mathrm{p}_t(\theta)} &= y_t^2 - \frac{2\, y_t}{\theta} \end{aligned}$$

and condition (3.10) is easily verified. For condition (3.11) observe that (using the third derivative)

$$D_t^{(2)}(\theta_1) - D_t^{(2)}(\theta_2) \leq (\theta_1 - \theta_2)\, y_t^2\, \epsilon^{-a\, y_t}\,(3 - a\, y_t),$$

which means that

$$\sup_{\theta_1, \theta_2 \in C} [(\theta_1 - \theta_2)^{-1}\, \{D_t^{(2)}(\theta_1) - D_t^{(2)}(\theta_2)\}] \leq y_t^2\, e^{-a\, y_t}\,(3 - a\, y_t),$$

and condition (3.11) can now be verified.

Now, for any specific estimator we have to verify condition C3. A SFS based on the maximum likelihood is efficient since the MLE estimator is consistent, $\{\mathbf{P}_\theta, C\}$-as $|\hat{\theta}_t - \tilde{\theta}_t|^2 = O(t^{-(1+\epsilon)})$ for some $e > 0$, and the posterior moments are of the appropriate order (Ibragimov and Hasminskii, 1980). □

3.7 Counterexamples

In previous sections we presented sufficient conditions for the efficiency of a plug-in SFS, and we tried to relate the estimative properties of an estimator to the efficiency of the SFS it produces. In this section we present some examples which show that plug-in SFS's can be inefficient.

The first example shows that even in the case of independent identically distributed observations a SFS based on a Fisher efficient estimator can be inefficient, and therefore prequential efficiency is a stronger property than Fisher efficiency. The same example shows why in the χ^2 approach we needed condition C3 (especially (3.12)), and in the Kullback-Leibler approach we required the second term in the expected risk of the estimator

$$E_\pi E_\theta\{K_{t+1}(\theta, \hat{\theta}_t)\} < \frac{p}{2t} + b_t,$$

to be such that $\sum_t b_t < \infty$.

Example 3.7 Let (Y_t) be independent identically distributed Normal observations with unknown mean θ and known variance σ^2. As was shown in example (3.4) the plug-in SFS based on the sample mean $\hat{\theta}_t := \sum_{j=1}^t Y_j/t$ is efficient. Let $e_t = \hat{\theta}_t + 1/(\sqrt{t\log(t)})$. Then for every θ

$$E_\theta(\theta - \hat{\theta}_t)^2 = \frac{\sigma^2}{t}$$

and

$$E_\theta(\theta - e_t)^2 = \frac{\sigma^2}{t} + \frac{1}{t \cdot \log(t)}.$$

The estimator e_t is asymptotically efficient since $t \cdot E_\theta(\theta - e_t)^2$ converges to σ^2. Let $\mathbf{Q}$ be the SFS based on $\hat{\theta}_t$, and $\mathbf{R}$ the SFS based on e_t. Under $\mathbf{Q}$ and $\mathbf{R}$, $Y = (Y_1, Y_2, ...)$ is a Gaussian process, and therefore a necessary condition for $\mathbf{Q}$ and $\mathbf{R}$ to be equivalent (Shiryayev, 1996, page 533) is that $\{\mathbf{P}_\theta, \Theta\}$-as

$$\sum_{t=1}^{\infty} \frac{(\hat{\theta}_t - e_t)^2}{\sigma^2} < \infty.$$

Simple calculations show that

$$\sum_{t=1}^{\infty} \frac{(\hat{\theta}_t - e_t)^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{t=1}^{\infty} \frac{1}{t \cdot \log(t)} = \infty,$$

and therefore the SFS $\mathbf{R}$ is inefficient.

In this example the estimator $\tilde{\theta}_t$ is first order efficient, but the extra term $1/\sqrt{t \log(t)}$ introduces an inefficiency of order $1/(t \cdot \log(t))$ which, although it does not affect the first order asymptotics in estimation terms, does affect the prequential efficiency of the plug-in SFS.

The next example shows that there are cases where, although the support of the forecast distributions does not depend on the parameter θ, there are no efficient plug-in SFS. This is because the Fisher information of every new observation is large with respect to the Fisher information of the previous data.

Example 3.8 Let (Y_t) be independent Normal observations with unknown mean θ and variance σ_t^2, known and positive for all t. Let the prior for θ be a Normal distribution with mean 0 and variance 1. Then the Bayesian predictive distribution for the observation Y_{t+1}, given Y^t, is a normal distribution with variance V_{t+1}:

$$V_{t+1} = \sigma_{t+1}^2 + \frac{1}{1 + S_t}$$

where $S_t = \sum_{j=1}^{t} 1/\sigma_j^2$. This yields an efficient SFS. The predictive distribution of a plug-in system will be Normal with variance σ_{t+1}^2, since the variance is known. Under both the BFS and any plug-in SFS the sequence $Y = (Y_1, Y_2, ...)$ is a Gaussian process, but of course with different means and variances. A necessary condition then for the plug-in SFS to be efficient is that (Shiryayev, 1996, page 533) $\{\mathbf{P}_\theta, \Theta\}$-as

$$\sum_{t=1}^{\infty} \left(\frac{V_{t+1}}{\sigma_{t+1}^2} - 1 \right)^2 < \infty.$$

But

$$\frac{V_{t+1}}{\sigma_{t+1}^2} - 1 = \frac{1}{(1 + S_t)\,\sigma_{t+1}^2},$$

and therefore if the variance σ^2_{t+1} is small with respect to $(1 + S_t)^{-1}$, then there exist no efficient plug-in SFS. For example this is the case when there is a constant c_1 such that for every t

$$\sigma^2_{t+1} < \frac{c_1 \sqrt{t}}{S_t}.$$

Actually, because under the two measures the process Y is a Gaussian process the two measures are then singular, and therefore $\{\mathbf{P}_\theta, \Theta\}$-as

$$\lim_{t \to \infty} \frac{d\mathbf{Q}^t}{d\mathbf{B}^t} = 0,$$

which means that any plug-in system is infinitely worse than the BFS.

We must note that in this example the MLE estimator satisfies almost every optimality criterion for estimation, but the fact that the uncertainty of $\hat{\theta}_t$ is not incorporated in the predictive distribution makes the plug-in SFS inefficient. □

Another example where no efficient plug-in forecasting systems exist is presented next.

Example 3.9 (Stochastic Linear Regression) Assume that the observations (Y_t) are generated from the following model:

$$Y_t = \theta' x_t + \epsilon_t, \tag{3.21}$$

where θ is an unknown vector of order p, the predictors x_t are fixed or predictable with respect to the filtration $\mathcal{F}_t = \sigma(Y_1, Y_2, \ldots, Y_t)$, and the errors ϵ_t are Normal with mean zero and variance σ^2, known and positive.

Let $\mathbf{B}$ be a BFS based on a Normal prior with mean zero and variance-covariance matrix the $p \times p$ identity matrix I_p. The predictive distribution of $\mathbf{B}$ for the observation Y_{t+1} is a Normal distribution with mean $\tilde{\theta}_t{}' x_{t+1}$ and variance $\sigma^2 \{1 + x'_{t+1} (X'_t X_t + I_p)^{-1} x_{t+1}\}$, where

$$\tilde{\theta}_t = (X'_t X_t + I_p)^{-1} X_t \mathbf{Y}_t,$$

$$X_t = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_t' \end{bmatrix}$$

and

$$\mathbf{Y}_t = (Y_1, Y_2, \ldots, Y_t)'.$$

Also let $\mathbf{Q}$ be the plug-in PFS based on the estimator $\tilde{\theta}_t$. The predictive distribution $\mathbf{Q}_{t+1}$ is also a Normal distribution with mean $\tilde{\theta}_t{}' x_{t+1}$, and variance equal to σ^2, since it is considered known.

The BFS $\mathbf{B}$ is efficient, and the plug-in PFS $\mathbf{Q}$ will be efficient if and only if $\{\mathbf{P}_\theta, \Theta\}$-as

$$\sum_{t=1}^{\infty} \{x_{t+1}'(X_t'X_t + I_p)^{-1} x_{t+1}\}^2 < \infty, \tag{3.22}$$

This condition follows from the fact that under $\mathbf{B}$ and $\mathbf{Q}$ the process Y is a Gaussian processes (Shiryayev, 1996). It can be shown that when (3.22) fails, there are no efficient plug-in SFS's. Note that when $x_{t+1}'(X_t'X_t)^{-1}x_{t+1}$ is bounded below, neither the BFS nor the plug-in SFS are consistent, in that $H(\mathbf{P}_{t+1,\theta}, \mathbf{B}_{t+1})$ and $H(\mathbf{P}_{t+1.\theta}, \mathbf{P}_{t+1.\hat{\theta}_t})$ do not converge to zero. Nevertheless, the BFS is efficient according to our definition : no other SFS can do any better.

If we apply the above result to the case of an autoregressive model of order one, i.e. when x_t is Y_{t-1}, it can be shown (Wei, 1987) that (3.22) holds only when $|\theta| \leq 1$. As a result, we see that there are no efficient plug-in SFS's for an explosive AR(1) model. When $|\theta| \leq 1$, a SFS based on the least squares estimator or any ridge estimator is efficient. □

3.8 Discussion

Our investigation showed that a class of non-Bayesian SFS's, the plug-in SFS's, are efficient, under suitable regularity conditions, and can thus be used either for pre-

diction or model selection. The main advantage of these SFS's is that in most cases they are easy to use, as well as being good approximations to "better" forecasting systems.

We also showed that a plug-in SFS should not be used without some investigation of the properties of the parametric model, since, as the counterexamples show, it could behave very badly, especially in cases where the Fisher information of the next observation is large with respect to the information for the data at hand. In such cases, while a Bayesian forecasting system will produce good forecasts, a plug-in SFS will be heavily penalised for the fact that it does not incorporate the uncertainty for the estimator of θ in its predictive distribution. In these cases a modification of the plug-in SFS may improve it. For example, instead of a plug-in predictive distribution we might use any of the predictive distributions proposed by Harris (1989),El-Sayyad et al. (1989), Kuboki (1993) and Basu and Harris (1994). The efficiency of such systems remains to be investigated.

The definition of prequential efficiency is based on the predictive assessment of a forecasting system using the logarithmic score. Can we extend this property to other forms of prediction and loss functions? A first step in this direction is described in the next chapter, where a similar property of efficiency is defined and studied for point prediction under squared error loss.

Chapter 4

Efficient Point Prediction Systems

4.1 Introduction

Our main objective in this chapter is to extend the notion of prequential efficiency for probability forecasting systems to point prediction, and to study the efficiency of different methods of constructing predictions. Specifically, we show that, under weak conditions, Bayesian predictors are efficient.

In a decision-theoretic framework an optimal point predictor is defined as that which minimises the expected loss. For example, if we observe X and want to predict Y under squared-error loss, then the predictor $g(X)$ that minimizes $E\{Y - g(X)\}^2$ is the conditional mean $E(Y|X)$. This is called the minimum mean squared error predictor of Y given X.

The problem with the above definition is that it presupposes knowledge of the joint distribution of X and Y. In most situations we do not have this information. Suppose instead that we can assume that the joint distribution of (X, Y) belongs to a parametric family of distributions, indexed by a parameter θ. In this case we

typically cannot use the above optimal predictor, since for every θ we will have a different optimal predictor $E(Y|X,\theta)$. There are now two sources of uncertainty, the predictand Y and the unknown parameter θ.

One simple and common way to proceed is to replace the unknown parameter θ in the optimal predictor with a suitable estimate of θ based on the observation of X. This is the plug-in approach, similar to the plug-in method for probability forecasting discussed in Chapter 3, and usually this method gives reasonable point predictions.

There is also a Bayesian method of issuing point predictions. We specify a prior distribution on the set of values for θ, thus completing the joint distribution of (θ, X, Y). Given a suitable loss function for prediction, the optimal predictor, from a decision-theoretic point of view, is that minimising the Bayes risk. For example, when the loss function is squared prediction error, the Bayes predictor is the function $g(X)$ which minimises the overall expectation $E\{Y - g(X)\}^2$, *viz.* $E(Y|X)$, where the expectations are calculated in the joint distribution of (X, Y) after marginalising over the random variable θ.

However, this method is not likely to be acceptable to a non-Bayesian statistician unless it can be shown to have good properties under the true model. Such properties are often phrased in terms of expectations conditional on θ, but these in turn might be objectionable to the Bayesian. In order to side-step such controversies, we introduce a new notion of optimality, in an asymptotic sequential framework. Our definition is based on the actual empirical performance of the prediction rule, and avoids references to conceptual replications of the setup for its justification. It is perhaps a disadvantage that our definition is based on infinite sequences and asymptotic arguments. Nevertheless, it can be used to study the actual performance of standard methods of constructing predictors, and help us discuss issues of optimality from a different perspective.

In this chapter we investigate this new approach for point predictors assessed

by squared prediction error loss, and we show that, under weak conditions, there exists a class of optimal predictors which we term *efficient*.

In §4.2 we describe the setup, introduce the notion of a *point prediction system (PPS)* as a rule that generates predictions sequentially, and present our definition of efficiency for PPS's. In §4.3 we study the case when the true distribution is known. This will provide us with the tools to show in §4.4 that efficient PPS's exist for general parametric families, by establishing that efficient PPS's can be constructed using a Bayesian approach. In §4.5 we discuss the relationship between efficient point prediction systems, and efficient probability forecasting systems. In §4.6 we present sufficient conditions for the efficiency of plug-in PPS's. In §4.7 we discuss some applications in probability forecasting and stochastic regression models.

4.2 Efficiency of Point Prediction Systems

We use a framework similar to the ones used in the previous chapters. Assume that a forecaster observes a sequence of random vectors $Y = (Y_t)$, $t \geq 1$, with $Y_t \in \mathbb{R}^k$. His task at each step t is to issue a point prediction for the next observation Y_{t+1} using the past observations $Y^t := (Y_1, \ldots, Y_t)$. The dimension k could change with t, but for simplicity we consider it fixed.

In order to issue his predictions the forecaster uses a rule, which for every set of outcomes for Y^t, and any other external information he may have at that time, specifies a point forecast for Y_{t+1}. We call such a rule a *Point Prediction System (PPS)*. The class of all PPS's is extremely broad, including any method of constructing one step ahead predictions.

To be more rigorous, let the sequence of the observed variables $Y = (Y_t)$ be defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbf{P})$. The filtration $(\mathcal{F}_t)$ represents the information available to the forecaster at each time point t. When the only information available at time t is the past observations Y^t, then $\mathcal{F}_t$ is the σ-field

generated by Y^t.

Following standard terminology (Shiryayev, 1996), we call a sequence of random elements (e.g. variables, vectors or matrices), $U = (U_t)$, a *stochastic* sequence if, for every t, U_t is $\mathcal{F}_t$-measurable; and *predictable* if each U_t is $\mathcal{F}_{t-1}$-measurable. We assume that the sequence of observables (Y_t) is a stochastic sequence of vectors in $\mathbb{R}^k$. Since any point prediction for Y_{t+1} should be based only on the information available to the forecaster at time t, any PPS is equivalent to a predictable sequence of vectors in $\mathbb{R}^k$, and vice-versa.

If A is a PPS, and $(A_1, A_2, \ldots, A_T)$ are the predictions it issues for the first T observations $(Y_1, Y_2, \ldots, Y_T)$, then the empirical performance of A up to time T may be assessed by the sum of the squared prediction errors:

$$S_T(A) = \sum_{t=1}^{T} \|Y_t - A_t\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^k$. This criterion also covers the seemingly more general case where we want to assess the performance of a PPS using weighted squared prediction errors

$$S_T^W(A) = \sum_{t=1}^{T} (Y_t - A_t)' W_t (Y_t - A_t),$$

where $W = (W_t)$ is a sequence of predictable symmetric positive definite matrices. This is because we can make the transformations $Y_{W,t} = W_t^{1/2} Y_t$ and $A_{W,t} = W_t^{1/2} A_t$, and use the criterion $S_T(\cdot)$ on the transformed variables and PPS.

In the light of any sequence of data, we can compare two PPS's, say A and D, using the difference between the cumulative loss for A, $S_T(A)$, and that for D, $S_T(D)$:

$$D_T(A, D) = S_T(A) - S_T(D).$$

For any PPS A, $S_T(A)$ is increasing in T, and typically tends to infinity. If for a specific infinite sequence of outcomes the difference $D_T(A, D)$ tends to $-\infty$, then we can consider that the PPS A has performed better than D, and the opposite if

$D_T(A, D)$ tends to $+\infty$. When the difference $D_T(A, D)$ stays bounded both above and below, we cannot effectively distinguish between the two PPS's, which can then be considered equivalent.

The above consideration motivates the following definition of an *efficient* PPS.

Definition 4.1 *Let $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ be a parametric family of probability measures on $(\Omega, \mathcal{F}, (\mathcal{F}_t))$. A PPS A will be called* efficient *for the family $\mathcal{P}$ if, for any other PPS D,*

$$\limsup_{T-\infty} D_T(A, D) < \infty, \tag{4.1}$$

*with $\mathbf{P}_\theta$-probability one for almost all θ in Θ (*i.e. *excepting perhaps a set of Lebesgue-measure zero).*

When Θ is countable, the same definition can be used, on replacing Lebesgue measure with counting measure (and thus rendering unnecessary the qualification "almost all").

According to our definition, a PPS is efficient if, with probability one for almost all θ, its empirical predictive performance will be at least as good as that of any other PPS. This is a strong property, and indeed when Θ is uncountable we cannot expect to find a PPS for which property (4.1) holds for *all* $\theta \in \Theta$. This is because it is typically easy to construct a PPS which is exceptionally good for some specific values or values of θ (*e.g.* based on assuming some particular value to be that generating the data). This phenomenon is akin to that of "super-efficiency". We shall see that, with the extra qualification "almost all", this difficulty can be avoided.

Our notion of efficiency is similar to the notion of *prequential efficiency* for probability forecasting systems introduced by Dawid (1984), and discussed in Chapters 2 and 3. Both are based on the idea that a prediction rule should be assessed by its empirical performance for the actual data arising. An efficient prediction system is then one that can be almost guaranteed to deliver optimal performance, so long as the data sequence arises from some probability distribution in the family

considered. With the important exception of the interpretation of "almost" in this statement, these definitions make no references to hypothetical replications of the setup, alternative data sequences, or events that did not occur. In particular, no concept of *expected* performance has any rôle to play.

4.3 Known Probability Distribution

In this preliminary section we assume that the probability measure $\mathbf{P}$ is fully known, and for every t the conditional means

$$M_t := E_{\mathbf{P},t-1}(Y_t),$$

and covariance matrices

$$S_t := E_{\mathbf{P},t-1}\left\{(Y_t - M_t)(Y_t - M_t)'\right\},$$

are defined $\mathbf{P}$-a.s., where $E_{\mathbf{P},t}(\cdot)$ denotes the conditional expectation $E_{\mathbf{P}}(\cdot|\mathcal{F}_t)$. For every t, S_t is a non-negative definite matrix. We denote its maximum eigenvalue by $\lambda \max S_t$. Since the distribution of $Y = (Y_1, Y_2, \ldots)$ is known, a PPS is efficient if property (4.1) holds with probability one under $\mathbf{P}$.

Although we assume that the conditional means M_t exist $\mathbf{P}$-a.s., we do not make any assumptions on the overall expectation of Y_t, and all the conditional expectations we use are generalized conditional expectations, as defined in Shiryayev (1996, chapter 7). Thus the stochastic sequence $U_T = \sum_{t=1}^{T}(Y_t - M_t)$ is a generalized martingale (Shiryayev, 1996, page 476), but not necessarily a martingale.

The sequence of the conditional means $M = (M_t)$ is a predictable sequence of vectors, and hence a PPS. For every t, the prediction M_t minimises the one-step-ahead predictive risk, $E_{\mathbf{P},t-1}(\|Y_t - g\|^2)$ over all $\mathcal{F}_{t-1}$-measurable functions g, and consequently M specifies the optimal (in decision-theoretic terms) sequence of one-step-ahead predictions. The next theorem studies the asymptotic performance

of the PPS M with respect to any other PPS. For two events E_1 and E_2 we say that **P**-a.s. $E_1 \Rightarrow E_2$ if the event that E_1 holds but E_2 fails has **P**-probability zero.

Theorem 4.1 *For any PPS $A = (A_t)$, **P**-a.s.*

$$\{\sup_t \lambda \max S_t < \infty\} \Rightarrow \{\lim_{T\to\infty} D_T(M,A) \text{ exists and is less than } \infty\}.$$

*More specifically, **P**-a.s.*

$$\left\{\sup_t \lambda \max S_t < \infty \text{ and } \sum_{t=1}^{\infty} \|M_t - A_t\|^2 < \infty\right\} \Rightarrow \left\{-\infty < \lim_{T\to\infty} D_T(M,A) < \infty\right\},$$

and

$$\left\{\sup_t \lambda \max S_t < \infty \text{ and } \sum_{t=1}^{\infty} \|M_t - A_t\|^2 = \infty\right\} \Rightarrow \left\{\lim_{T\to\infty} D_T(M,A) = -\infty\right\}.$$

Proof of Theorem 4.1. First we prove the following lemma:

Lemma 4.1 *If S_t is a generalized martingale then S_t^2 is a generalized submartingale, and if A_t is the compensator of S_t^2 then **P**-a.s.*

$$A_\infty < \infty \Longrightarrow S_t \text{ converges to a finite limit}$$

Proof. This proof is a simple adaptation of Theorem 3 in page 518 of Shiryayev (1996) for generalized submartingales, but we present it for completeness. The stochastic sequence $(S_t+1)^2$ is a nonnegative generalized submartingale with compensator $A_t + 1$. We know (Shiryayev, 1996, page 523) that for a nonnegative generalized submartingale if the limit of the compensator is finite then **P**-a.s. the submartingale converges to a finite limit. Therefore S_t^2 and $(S_t+1)^2$ converge to finite limits, and therefore S_t converges to a finite limit since

$$S_t = \frac{1}{2}\left\{(S_t+1)^2 - S_t^2 - 1\right\}.$$

Proof of Theorem. It is easy to show that

$$E_{\mathbf{P},t-1}(\|Y_t - A_t\|^2) = E_{\mathbf{P},t-1}(\|Y_t - M_t\|^2) + \|M_t - A_t\|^2.$$

The stochastic sequence (Δ_T), where

$$\begin{aligned}\Delta_T &= D_T(M,A) + \sum_{t=1}^{T} \|M_t - A_t\|^2 \\ &= 2\sum_{t=1}^{T}(A_t - M_t)'(Y_t - M_t),\end{aligned}$$

is a generalized martingale.

Case 1. $\sup_t \lambda \max S_t < \infty$ and $\sum_{t=1}^{\infty} \|A_t - M_t\|^2 < \infty$.

Since

$$E_{\mathbf{P},T}(\Delta_{T+1}^2) = \Delta_T^2 + 4\,(A_{T+1} - M_{T+1})' S_{T+1}(A_{T+1} - M_{T+1}),$$

the sequence (Δ_T^2) is a generalized submartingale, with compensator $C_T = 4\sum_{t=1}^{T}(A_t - M_t)' S_t(A_t - M_t)$ which can be bounded above:

$$C_T \le 4 \max_{t \le T} \lambda \max S_t \sum_{t=1}^{T} \|A_t - M_t\|^2.$$

It follows that if $\sum_{t=1}^{\infty} \|A_t - M_t\|^2 < \infty$, then C_T is finite, and using Lemma (4.1), Δ_T converges to a finite limit **P**-a.s.. The stochastic sequence $D_T(M,A)$ also converges **P**-a.s. to a finite limit since $D_T(M,A) = \Delta_T - \sum_{t=1}^{T} \|A_t - M_t\|^2$.

Case 2. $\sup_t \lambda \max S_t < \infty$ and $\sum_{t=1}^{\infty} \|A_t - M_t\|^2 = \infty$.

Let $a_T = \max(1, \sum_{t=1}^{T} \|A_t - M_t\|^2)$. Consider the generalized martingale

$$W_T = \sum_{t=1}^{T} \frac{2\,(A_t - M_t)'(Y_t - M_t)}{a_t}.$$

Then W_T^2 is a nonnegative generalized submartingale with compensator

$$C_T = 4\sum_{t=1}^{T} \frac{(A_t - M_t)' S_t (A_t - M_t)}{a_t^2},$$

which is bounded above by

$$4\left\{\max_{t \le T} \lambda \max S_t\right\} \sum_{t=1}^{T} \frac{\|A_t - M_t\|^2}{a_t^2}.$$

Using Lemma 4.1 we see that the sequence (W_T) converges **P**-a.s. to a finite limit. Since

$$\frac{\Delta_T}{a_T} = \frac{\sum_{t=1}^{T} a_t(W_t - W_{t-1})}{a_T},$$

from Kronecker's lemma, **P**-a.s.

$$\lim_{T\to\infty} \frac{\Delta_T}{a_T} = 0,$$

and

$$\lim_{T\to\infty} \frac{D_T(M,A)}{a_T} = -1.$$

It follows that $\lim_{T\to\infty} D_T(M,A) = -\infty$, and the theorem has been established. □

Theorem 4.1 shows that when the covariance matrices S_t stay bounded in all directions **P**-a.s., then the PPS M is efficient. Any other PPS A is also efficient if and only if it issues predictions which are asymptotically equivalent to the predictions issued by the PPS M, i.e. if and only if **P**-a.s.

$$\sum_{t=1}^{\infty} \|M_t - A_t\|^2 < \infty.$$

Note that any two efficient PPS's, say $A = (A_t)$ and $D = (D_t)$, asymptotically issue equivalent predictions not only for the next observation, but for the whole infinite future, in the sense that, **P**-a.s.,

$$\lim_{T\to\infty} \sum_{t=T}^{\infty} \|A_t - D_t\|^2 = 0,$$

which holds since

$$\sum_{t=T}^{\infty} \|A_t - D_t\|^2 \le 2\left(\sum_{t=T}^{\infty} \|M_t - A_t\|^2 + \sum_{t=T}^{\infty} \|M_t - D_t\|^2\right).$$

The next example shows that, in order to establish the above results, we do need to bound somehow the increase of the covariances. Otherwise we can get surprising results: for example, there exist cases where efficient systems do exist, but M is not one of them.

Example 4.1 Assume that we observe a sequence of independent observations (Y_t), with $\mathbf{P}(Y_t = t^2) = 1/t^2$ and $\mathbf{P}(Y_t = 0) = 1 - 1/t^2$. Then, for every t, $M_t = E(Y_t) = 1$ and according to the PPS M the best prediction for each Y_t is 1. But noting that

$$\sum_{t=1}^{\infty} \mathbf{P}(Y_t \neq 0) = \sum_{t=1}^{\infty} \frac{1}{t^2} < \infty,$$

we have, from the Borel-Cantelli lemma, that $\mathbf{P}$-a.s. there exists t_0 (depending on the outcome sequence) such that, for every $t > t_0$, $Y_t = 0$. Therefore the best sequence of predictions in this case comes from a PPS, say $Z = (Z_t)$, which for every t predicts 0 for Y_t. With probability one, the loss $S_T(Z)$ stays finite as T tends to infinity, and any other PPS $A = (A_t)$, is efficient if and only if $\mathbf{P}$-a.s.

$$\sum_{t=1}^{\infty} \|A_t\|^2 < \infty.$$

It follows that the PPS M is not efficient, since

$$\sum_{t=1}^{\infty} \|M_t\|^2 = \sum_{t=1}^{\infty} 1 = \infty.$$

□

In view of the above example we see that, without further conditions, optimality defined in terms of the minimization of the one-step ahead risk does not necessarily imply optimality in terms of the asymptotic empirical performance of a prediction rule.

4.4 Parametric Family of Distributions

Assume now that $\mathbf{P}$ is an unknown member of a parametric family of probability measures $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$ on $(\Omega, \mathcal{F}, (\mathcal{F}_t))$, where Θ is a subset of $\mathbb{R}^p$, $p \geq 1$. For an event A, and subset C of Θ, we say that A holds $\{\mathbf{P}_\theta, C\}$-as if $\mathbf{P}_\theta(A) = 1$ for almost all θ in C (*i.e.* excepting perhaps a set of Lebesgue measure zero). We remind that for any probability measure $\mathbf{Q}$ on $(\Omega, \mathcal{F}, (\mathcal{F}_t))$, we denote by $E_{\mathbf{Q},t}(\cdot)$

the conditional expectation $E_{\mathbf{Q}}(\cdot|\mathcal{F}_t)$, and by $m_t(\theta)$ the conditional mean of Y_t given $\mathcal{F}_{t-1}$ under $\mathbf{P}_\theta$, *i.e.*

$$m_t(\theta) := E_\theta(Y_t|\mathcal{F}_{t-1}) := E_{\theta,t-1}(Y_t).$$

We now show how efficient PPS's can be constructed using a Bayesian approach. Let $\pi(\theta)$ be a prior probability density for θ, which is almost everywhere positive on Θ. Then we can define the Bayesian marginal measure on $(\Omega, \mathcal{F}, (\mathcal{F}_t))$:

$$\mathbf{B} = \int_\Theta \mathbf{P}_\theta \, \pi(\theta) \, d\theta. \tag{4.2}$$

At time t, we can calculate the posterior density $\pi_t(\theta)$ of θ given $\mathcal{F}_t$. If we assume that $\{\mathbf{P}_\theta, \Theta\}$-as the conditional means

$$M_{\mathbf{B},t} = E_{\mathbf{B},t-1}(Y_t)$$

and covariances

$$S_{\mathbf{B},t} = E_{\mathbf{B},t-1}\{(Y_t - M_{\mathbf{B},t})(Y_t - M_{\mathbf{B},t})'\}$$

are finite for every t. and

$$\sup_t \lambda \max S_{\mathbf{B},t} < \infty, \ \{\mathbf{P}_\theta, \Theta\}\text{-as}, \tag{4.3}$$

then we can prove the following theorem.

Theorem 4.2 *Let* $\mathbf{B}$ *be defined as in* (4.2), *and assume that it satisfies assumption* (4.3). *If* $M_{\mathbf{B}} := (M_{\mathbf{B},t})$ *is the Bayesian PPS (BPPS) based on the conditional means of* $\mathbf{B}$, *and* $A = (A_t)$ *is any other PPS, then* $\{\mathbf{P}_\theta, \Theta\}$*-as*

$$\lim_{T\to\infty} D_T(M_{\mathbf{B}}, A) < \infty.$$

More specifically, $\{\mathbf{P}_\theta, \Theta\}$*-as*

$$\{\sum_{t=1}^{\infty} \|M_{\mathbf{B},t} - A_t\|^2 < \infty\} \Rightarrow \{-\infty < \lim_{T\to\infty} D_T(M_{\mathbf{B}}, A) < -\infty\},$$

and

$$\{\sum_{t=1}^{\infty} \|M_{\mathbf{B},t} - A_t\|^2 = \infty\} \Rightarrow \{\lim_{T\to\infty} D_T(M_{\mathbf{B}}, A) = -\infty\}.$$

Proof of Theorem 4.2. Using Theorem 4.1 we know that, for any other PPS $A = (A_t)$,

$$\{\sup_t \lambda \max S_{\mathbf{B},t} < \infty\} \Rightarrow \{\lim_{T\to\infty} D_T(M_{\mathbf{B}}, A) < \infty\}, \ \mathbf{B}\text{-a.s.} \tag{4.4}$$

If an event has probability one under $\mathbf{B}$, then it has probability one for all θ in Θ, except perhaps for a set of Lebesgue measure zero, since the prior density $\pi(\theta)$ is almost everywhere positive on Θ. Therefore (4.4) holds $\{\mathbf{P}_\theta, \Theta\}$-as. From assumption (4.3), the event

$$\sup_t \lambda \max S_{\mathbf{B},t} < \infty$$

holds $\{\mathbf{P}_\theta, \Theta\}$-as and therefore $\{\mathbf{P}_\theta, \Theta\}$-as

$$\lim_{T-\infty} D_T(M_{\mathbf{B}}, A) < \infty,$$

and $\{\mathbf{P}_\theta, \Theta\}$-as

$$\{\sup_t \lambda \max S_{\mathbf{B},t} < \infty \text{ and } \sum_{t=1}^{\infty} \|M_{\mathbf{B},t} - A_t\|^2 < \infty\} \ \Rightarrow \ \{-\infty < \lim_{T\to\infty} D_T(M_{\mathbf{B}}, A) < -\infty\},$$

$$\{\sup_t \lambda \max S_{\mathbf{B},t} < \infty \text{ and } \sum_{t=1}^{\infty} \|M_{\mathbf{B},t} - A_t\|^2 = \infty\} \ \Rightarrow \ \{\lim_{T\to\infty} D_T(M_{\mathbf{B}}, A) = -\infty\}.$$

□

Thus, as the last theorem shows, a PPS based on the conditional means of a Bayesian measure is efficient under the minimal assumption that its predictive covariance matrices (calculated under the marginal measure $\mathbf{B}$) stay bounded with probability one for almost all θ. It is easy to show that

$$M_{\mathbf{B},t} = \int_\Theta m_t(\theta)\ \pi_{t-1}(\theta)\, d\theta,$$

and therefore the Bayesian prediction is a weighted average of the conditional means $m_t(\theta)$, where the posterior density $\pi_{t-1}(\theta)$ of θ given $\mathcal{F}_{t-1}$ provides the weights for the different θ's.

For an arbitrary PPS $D = (D_t)$ we can establish its efficiency by comparing its predictions with those of an efficient BPPS, and by using the fact that D is efficient if and only if asymptotically the predictions of the two PPS's are equivalent, *i.e.* if

$$\sum_{t=1}^{\infty} \|M_{\mathbf{B},t} - D_t\|^2 < \infty, \ \{\mathbf{P}_\theta, \Theta\}\text{-as.} \tag{4.5}$$

Using the submartingale convergence theorem it can be shown that a sufficient condition for (4.5) to hold is that

$$\sup_T \sum_{t=1}^{T} E_{\mathbf{B}}(\|M_{\mathbf{B},t} - D_t\|^2) < \infty.$$

A typical expectation in the above sum can be written as

$$E_{\mathbf{B}}\|M_{\mathbf{B},t} - D_t\|^2 = E_\pi E_\theta \|m_t(\theta) - D_t\|^2 - E_\pi E_\theta \|m_t(\theta) - M_{\mathbf{B},t}\|^2,$$

where, for every term, the first expectation is with respect to the prior density $\pi(\theta)$, and the second with respect to the probability measure $\mathbf{P}_\theta$. Thus the expectation $E_{\mathbf{B}}\|M_{\mathbf{B},t} - D_t\|^2$ is equal to the difference between the Bayes risks of $M_{\mathbf{B},t}$ and D_t for the estimation of the conditional mean $m_t(\theta)$ under squared error loss. The prediction $M_{\mathbf{B},t}$ minimizes this risk, and hence a sufficient condition for the efficiency of the PPS D is that its cumulative Bayes risk be sufficiently close to the minimum, achieved by the Bayesian PPS.

For the case where the probability measure $\mathbf{P}$ was known, we were able to show that any two efficient PPS's issue asymptotically equivalent predictions not only for the next observation, but for the infinite future. This result can be extended to the case of a parametric family.

Theorem 4.3 *Let* $\mathbf{B}$ *be a Bayesian measure, and assume that* (4.3) *holds. Let* $D = (D_t)$ *and* $A = (A_t)$ *be any two efficient PPS for the family* $\mathcal{P}$. *Then* $\{\mathbf{P}_\theta, \Theta\}$*-as*

$$\lim_{T\to\infty} \sum_{t=T}^{\infty} \|D_t - A_t\|^2 = 0.$$

Proof of Theorem 4.3. Let $M_{\mathbf{B}} = (M_{\mathbf{B},t})$ be the BPPS based on the conditional means of $\mathbf{B}$. Then using Theorem 4.2, $M_{\mathbf{B}}$ is efficient and the PPS's $D = (D_t)$ and $A = (A_t)$ are efficient if and only if $\{\mathbf{P}_\theta, \Theta\}$-as

$$\sum_{t=1}^{\infty} \|M_{\mathbf{B},t} - D_t\|^2 < \infty$$

and

$$\sum_{t=1}^{\infty} \|M_{\mathbf{B},t} - A_t\|^2 < \infty.$$

Clearly then $\{\mathbf{P}_\theta, \Theta\}$-as

$$\sum_{t=1}^{\infty} \|D_t - A_t\|^2 < \infty,$$

and finally $\{\mathbf{P}_\theta, \Theta\}$-as

$$\lim_{T-\infty} \sum_{t=T}^{\infty} \|D_t - A_t\| = 0.$$

□

4.5 Efficient SFS and Point Prediction

The notion of a *statistical forecasting system* (SFS) was introduced in section 2.4. For completeness we repeat here that a SFS is a prediction rule which, for every t and every realisation of the outcome of $\mathcal{F}_t$, specifies a predictive distribution for the next observation Y_{t+1}. Any SFS is consistent with at least one joint distribution $\mathbf{Q}$ on $(\Omega, \mathcal{F}, (\mathcal{F}_t))$. The property of *(prequential) efficiency* for probability forecasting systems was defined in section 2.4, in the simplest context with $\mathcal{F}_t = \sigma\{Y_1, \ldots, Y_t\}$. When the distribution $\mathbf{Q}$ is uniquely determined and may thus be equated with the SFS, this is as follows:

Definition 4.2 *Let* $\mathbf{Q}$ *be a SFS, and* $\mathbf{Q}^t$ *its restriction to* $\mathcal{F}_t$. *Then* $\mathbf{Q}$ *is termed* efficient *if, for any other PFS* $\mathbf{S}$, *the Radon-Nikodym derivative of* $\mathbf{S}^t$ *with respect to* $\mathbf{Q}^t$:

$$\Lambda_t(\mathbf{S}, \mathbf{Q}) := \frac{d\mathbf{S}^t}{d\mathbf{Q}^t},$$

converges to a finite limit with probability one for almost all θ in Θ.

In Seillier-Moiseiwitsch, Sweeting, and Dawid (1992) it was shown that any Bayesian probability forecasting system based on an almost everywhere positive prior is efficient, and any other PFS $\mathbf{Q}$ is efficient if and only if there is a BFS $\mathbf{B}$ such that $\mathbf{B} \ll \mathbf{Q}$. The result extends to cases where $\sigma\{Y_1, \ldots, Y_t\} \subset \mathcal{F}_t$.

In the previous section it was established, under a weak condition, that a PPS based on a Bayesian SFS is efficient. It is natural then to ask if the same result can be extended to any efficient probability forecasting system. The following theorem shows that this is possible, under a similar weak condition on the predictive variances of the efficient probability forecasting system.

Theorem 4.4 *Let $\mathbf{Q}$ be a probability measure on $(\Omega, \mathcal{F}, (\mathcal{F}_t))$, and assume that there exists a Bayesian PFS $\mathbf{B}$ such that $\mathbf{B} \ll \mathbf{Q}$. Let $S_{\mathbf{Q},t}$ denote the conditional covariance matrix of Y_t under $\mathbf{Q}$. Then if $\{\mathbf{P}_\theta, \Theta\}$-as*

$$\sup_t \lambda \max S_{\mathbf{Q},t} < \infty,$$

then the PPS based on the conditional means of (Y_t) under $\mathbf{Q}$ is efficient.

Proof of Theorem 4.4. Since $\mathbf{B} \ll \mathbf{Q}$, then any event that has probability one under $\mathbf{Q}$ has probability one under $\mathbf{B}$. The proof continues in the same way as the proof of Theorem 4.2. □

This theorem shows that one way of constructing an efficient PPS for parametric families is by using efficient probability forecasting systems. The only assumption that we have to check is whether the predictive covariance matrices stay bounded in all directions.

4.6 Plug-in PPS's

A popular method of forming predictions is by replacing the unknown parameter θ in the predictive mean $m_{t+1}(\theta)$, with an estimate $\hat{\theta}_t$ based on the available data

at time t. Then the prediction for Y_{t+1} is $m_{t+1}(\hat{\theta}_t)$. We will call a PPS based on this rule a *plug-in* PPS.

One method of establishing the efficiency of a plug-in PPS is by showing that it is generated by an efficient plug-in probability forecasting system, and verify the extra condition of Theorem 4.4. Sufficient conditions for the efficiency of a plug-in PFS are presented in Chapter 3.

Another more direct method is to compare the predictions of the plug-in PPS with the predictions of an efficient Bayesian PPS. Using this idea we next present sufficient conditions for the efficiency of a plug-in PPS. We assume that the plug-in PPS is based on some estimator sequence $(\hat{\theta}_t)$. The only property that the estimator sequence should satisfy is described in Condition 4. In order to have a well-defined plug-in PPS, we assume that $\hat{\theta}_t$ is defined for any $t \geq 0$, and takes values in Θ. This does not affect the generality of the results, since they are asymptotic.

Condition 1.

Let $\mathbf{B}$ be a Bayesian measure based on a prior probability density $\pi(\theta)$ for θ, almost everywhere positive, such that (4.3) holds. Let $\tilde{\theta}_t$ denote the posterior mean of θ based on the data $Y^t = (Y_1, Y_2, \ldots, Y_t)$, and V_t the posterior expectation of $\|\theta - \hat{\theta}_t\|^2$. Then $\{\mathbf{P}_\theta, \Theta\}$-as

$$\sum_{t=1}^{\infty} V_t^2 < \infty.$$

Condition 2.

The parameter set Θ is open, convex, and $\{\mathbf{P}_\theta, \Theta\}$-as the predictive means $m_t(\theta)$ are twice continuously differentiable with respect to θ, with derivatives $D^{(1)}m_t(\theta)$ and $D^{(2)}m_t(\theta)$.

Condition 3.

Let

$$d_t^{(1)}(\theta) = \|D^{(1)}m_t(\theta)\| \text{ and } d_t^{(2)}(\theta) = \sup_{\{h:h'h=1\}} |h'\, D^{(2)}m_t(\theta)\, h|.$$

Then $\{\mathbf{P}_\theta, \Theta\}$-as

$$\sup_t \sup_{\theta \in \Theta} \left\{ d_t^{(1)}(\theta) + d_t^{(2)}(\theta) \right\} < \infty.$$

Condition 4.

For the sequence of estimators $(\hat{\theta}_t)$, $\{\mathbf{P}_\theta, \Theta\}$-as

$$\sum_{t=1}^{\infty} \|\hat{\theta}_t - \tilde{\theta}_t\|^2 < \infty.$$

Theorem 4.5 *Assume that Conditions 1-4 hold. Then the plug-in PPS D based on the estimator sequence* $(\hat{\theta}_t)$ *is efficient.*

Comments:

(a) Condition 1 guarantees the existence of an efficient BPPS, which we use to establish the efficiency of the plug-in PPS. It also controls the behaviour of the posterior variance. This is a weak condition since typically the posterior variance is of order $O(t^{-1})$. The sum $\sum_{t=1}^{T} V_t^2$ is positive and increasing, and therefore a submartingale under any $\mathbf{P}_\theta$ or $\mathbf{B}$. A sufficient condition for the second part of Condition 1 to hold is that this sum stay finite in expectation for almost all θ, or under $\mathbf{B}$. This can be shown using the submartingale convergence theorem.

(b) Conditions 2 and 3 are smoothness conditions. Condition 3 is restrictive since it states that the derivatives stay bounded, but see comment (d).

(c) Condition 4 is the only condition that the estimator $\hat{\theta}_t$ should satisfy. Although this condition seems restrictive, it is almost a necessary condition as it is easy to construct examples where Condition 4 fails, and the PPS based on $\hat{\theta}_t$ is not efficient. For example assume that the observations (Y_t) are independent, having the Normal

distribution $N(\theta, 1)$. The PPS based on the sequence of the posterior means of θ for any Bayesian measure, based on a proper prior density, is efficient, and any plug-in PPS is efficient if and only if Condition 4 holds.

If $\hat{\theta}_t$ is a maximum likelihood estimator (MLE), then typically (when the observed information grows at rate t) its distance from the posterior mean is of order $O(1/t^{1/2+\epsilon})$ for some $\epsilon > 0$ (Johnson (1970), Ibragimov and Hasminskii (1980), Crowder (1988)). Therefore Condition 4 seems to be a weak condition for the MLE.

The sum of the squared distances between $\hat{\theta}_t$ and the posterior mean $\tilde{\theta}_t$ is an increasing positive sequence, and therefore a submartingale under any $\mathbf{P}_\theta$ and under $\mathbf{B}$. Using again the submartingale convergence theorem we have that a sufficient condition for Condition 4 to hold is that the sum stay finite in expectation for almost all θ or for $\mathbf{B}$. For example if

$$\sup_T E_{\mathbf{B}} \left(\sum_{t=1}^{T} \|\hat{\theta}_t - \tilde{\theta}_t\|^2 \right) < \infty, \tag{4.6}$$

then Condition 4 holds. Observe that

$$E_{\mathbf{B}} \|\hat{\theta}_t - \tilde{\theta}_t\|^2 = E_\pi E_\theta \|\hat{\theta}_t - \theta\|^2 - E_\pi E_\theta \|\tilde{\theta}_t - \theta\|^2.$$

The estimator $\tilde{\theta}_t$ minimizes the risk $E_\pi E_\theta \|e_t - \theta\|^2$ over all estimators e_t, and therefore in order to establish that Condition 4 holds using (4.6), it is sufficient to show that the estimative Bayes risk (under quadratic loss) of the estimator $\hat{\theta}_t$ is close to the minimum risk, achieved from the posterior mean $\tilde{\theta}_t$.

(*d*) Although in the statement of the theorem we used the whole parameter set Θ, the same assumptions and theorem can be used to show that a plug-in PPS is efficient for a subset C of Θ, using the results in section 2.5. Since Condition 3 involves suprema over the parameter set, in many situations it will not hold for the whole set Θ. What we can do then is to find a suitable countable cover of Θ, *i.e.*

a collection of subsets $(C_j)_{j\in\mathbb{N}}$ of Θ such that

$$\Theta = \bigcup_{j\in\mathbb{N}} C_j,$$

and to use the theorem for every one of the C_j's in order to show that the plug-in PPS is efficient for the subfamily of distributions $\mathbf{P}_{C_j} = \{\mathbf{P}_\theta : \theta \in C_j\}$ for every j. The subsets C_j should be chosen in a way such that Conditions 1–4 are satisfied for every one of them. Then, since it can be shown (Lemma 2.6) that if an event E is $\{\mathbf{P}_\theta, C_j\}$-as for every j, then it is $\{\mathbf{P}_\theta, \Theta\}$-as, the plug-in PPS is efficient for the whole family $\mathcal{P}$.

(*e*) It is possible that, while Conditions 1–4 do not hold in the original parametrisation, they might do so after a transformation of the parameter. Then the conclusion of the Theorem would hold, since it does not depend on the parametrisation used.

Proof of Theorem 4.5. Let $M_{\mathbf{B}}$ be the BPPS based on $\mathbf{B}$. Since the conditional variance matrices stay bounded, $M_{\mathbf{B}}$ is an efficient PPS. In order to show that D is efficient, it is sufficient to upper bound the squared distances between the predictions of $M_{\mathbf{B}}$ and D with suitable functions in order to establish that

$$\sum_{t=1}^{\infty} \|M_{\mathbf{B},t} - D_t\|^2 < \infty.$$

Using Condition 2, for every $\theta \in \Theta$ there is a θ^* (which may depend on θ) such that

$$\begin{aligned} \|m_{t+1}(\theta) - m_{t+1}(\hat{\theta}_t) - (\theta - \hat{\theta}_t)' D^{(1)}_{t+1}(\hat{\theta}_t)\| &= \|\tfrac{1}{2}(\theta - \hat{\theta}_t)' D^{(2)}_{t+1}(\theta^*)(\theta - \hat{\theta}_t)\| \\ &\leq \frac{1}{2}\|\theta - \hat{\theta}_t\|^2 \sup_{s\in\Theta} d^{(2)}_{t+1}(s) \end{aligned}$$

and therefore

$$\begin{aligned} \|M_{\mathbf{B},t+1} - m_{t+1}(\hat{\theta}_t)\| &\leq \|(\tilde{\theta}_t - \hat{\theta}_t)' D^{(1)}_{t+1}(\hat{\theta}_t)\| + \|M_{\mathbf{B},t+1} - m_{t+1}(\hat{\theta}_t) - (\tilde{\theta}_t - \hat{\theta}_t)' D^{(1)}_{t+1}(\hat{\theta}_t)\| \\ = \|(\tilde{\theta}_t - \hat{\theta}_t)' D^{(1)}_{t+1}(\hat{\theta}_t)\| &+ \|\textstyle\int_\Theta m_{t+1}(\theta)\,\pi_t(\theta)\,d\theta - m_{t+1}(\hat{\theta}_t) - (\tilde{\theta}_t - \hat{\theta}_t)' D^{(1)}_{t+1}(\hat{\theta}_t)\| \\ \leq \|\tilde{\theta}_t - \hat{\theta}_t\| \sup_{s\in\Theta} d^{(1)}_{t+1}(s) &+ \frac{1}{2}\left(V_t + \|\tilde{\theta}_t - \hat{\theta}_t\|^2\right) \sup_{s\in\Theta} d^{(2)}_{t+1}(s). \end{aligned}$$

Then

$$\begin{aligned}
\|M_{\mathbf{B},t+1} - m_{t+1}(\hat{\theta}_t)\|^2 &= \|\int_\Theta m_{t+1}(\theta)\,\pi_t(\theta)\,d\theta - m_{t+1}(\hat{\theta}_t)\|^2 \\
&\le 2\,\|\tilde{\theta}_t - \hat{\theta}_t\|^2 \left\{\sup_{s\in\Theta} d_{t+1}^{(1)}(s)\right\}^2 + \left(V_t + \|\tilde{\theta}_t - \hat{\theta}_t\|^2\right)^2 \left\{\sup_{s\in\Theta} d_{t+1}^{(2)}(s)\right\}^2 \\
&\le 2\,\|\tilde{\theta}_t - \hat{\theta}_t\|^2 \left\{\sup_{s\in\Theta} d_{t+1}^{(1)}(s)\right\}^2 + 2\,\|\tilde{\theta}_t - \hat{\theta}_t\|^4 \left\{\sup_{s\in\Theta} d_{t+1}^{(2)}(s)\right\}^2 \\
&\quad + 2\,V_t^2 \left\{\sup_{s\in\Theta} d_{t+1}^{(2)}(s)\right\}^2.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\sum_{t=0}^{T} \|\int_\Theta m_{t+1}(\theta)\,\pi_t(\theta)\,d\theta - m_{t+1}(\hat{\theta}_t)\|^2 \le \\
&\quad 2\left\{\sup_{t\le T}\sup_{s\in\Theta} d_{t+1}^{(1)}(s)\right\}^2 \sum_{t=0}^{T} \|\tilde{\theta}_t - \hat{\theta}_t\|^2 \\
&\quad + 2\left\{\sup_{t\le T}\sup_{s\in\Theta} d_{t+1}^{(2)}(s)\right\}^2 \sum_{t=0}^{T} \|\tilde{\theta}_t - \hat{\theta}_t\|^4 + 2\left\{\sup_{t\le T}\sup_{s\in\Theta} d_{t+1}^{(2)}(s)\right\}^2 \sum_{t=0}^{\infty} V_t^2.
\end{aligned}$$

By Condition 3, the quantities $\{\sup_{t\le T}\sup_{s\in\Theta} d_{t+1}^{(1)}(s)\}^2$ and $\{\sup_{t\le T}\sup_{s\in\Theta} d_{t+1}^{(2)}(s)\}^2$ stay bounded $\{\mathbf{P}_\theta, \Theta\}$-as. If we also use Conditions 1 and 4, it is clear that the three sums stay bounded $\{\mathbf{P}_\theta, \Theta\}$-as, as T tends to infinity. Then $\{\mathbf{P}_\theta, \Theta\}$-as the sum

$$\begin{aligned}
\sum_{t=0}^{\infty} \|M_{B,t+1} - D_{t+1}\|^2 &= \sum_{t=0}^{\infty} \|M_{B,t+1} - m_{t+1}(\hat{\theta}_t)\|^2 \\
&= \sum_{t=0}^{\infty} \|\int_\Theta m_{t+1}(\theta)\,\pi_t(\theta)\,d\theta - m_{t+1}(\hat{\theta}_t)\|^2
\end{aligned}$$

is finite. The proof is complete. □

The next example illustrates how Conditions 1–4 can be verified in specific examples.

Example 4.2 (Poisson Loglinear Model) Assume that Y_t are independent Poisson observations with means $\lambda_t = \exp(x_t\theta)$, where $x_t \in \mathbb{R}$ are fixed explanatory variables, and the unknown parameter θ takes values in $\mathbb{R}$. Then

$$m_t(\theta) = \exp(x_t\theta)$$

and therefore

$$d^{(1)}(\theta) = |x_t| \exp(x_t\theta)$$

and

$$d^{(2)}(\theta) = |x_t|^2 \exp(x_t\theta).$$

If we restrict our attention to a bounded subset $C = (\alpha, \beta)$ of Θ, then Conditions 2 and 3 hold if $\sup_t |x_t| < \infty$, since the parameter θ is bounded. Using the results in Crowder (1988) we can verify Conditions 1 and 4, for any Bayesian measure $\mathbf{B}$, if $\lim_{T-\infty} T^{-1} \sum_{t=1}^{T} x_t^2 > 0$. Therefore a PPS based on the maximum likelihood estimator is efficient for any subset (α, β) of Θ, and therefore efficient for the whole family since

$$\Theta = \bigcup_{j=1}^{\infty} (-j, j).$$

□

4.7 Applications

In this section we present some applications of our results to probability forecasting and stochastic regression.

Example 4.3 (Brier Score.) Let $E = (E_t)$ be a sequence of events of interest, and $D = (D_t)$ a PPS which issues a probability prediction D_t for E_t. For such a forecasting system the Brier score is defined as:

$$\mathrm{B}_T(D) = \sum_{t=1}^{T} \{I(E_t) - D_t\}^2,$$

where $I(\cdot)$ denotes indicator function. If $\mathbf{P}$ is the true probability measure, we denote by $M_{\mathbf{P}} = (M_{\mathbf{P},t})$ the PPS based on $\mathbf{P}$, with

$$M_{\mathbf{P},t} = E_{\mathbf{P}} \{I(E_t)|\mathcal{F}_{t-1}\} = \mathbf{P}(E_t|\mathcal{F}_{t-1}).$$

Applying Theorem 4.1 we have that $\mathbf{P}$-a.s.

$$\lim_{T\to\infty} \{\mathrm{B}_T(M_{\mathbf{P}}) - \mathrm{B}_T(D)\} < \infty,$$

and indeed **P**-a.s. each of the events

$$\sum_{t=1}^{\infty}(M_{\mathbf{P},t} - D_t)^2 < \infty$$

and

$$\lim_{T\to\infty} \{\mathrm{B}_T(M_{\mathbf{P}}) - \mathrm{B}_T(D)\} > -\infty$$

implies the other. This result means that the Brier score can be used as a consistent model selection rule for two distributions, since, with probability one, the true distribution **P** will eventually have a smaller Brier score than any other distribution **Q**, except in the case where the two distributions issue asymptotically equivalent forecasts.

When the true probability measure is not known, but belongs to a parametric family, we can show, using the results in §3 and §4, that any Bayesian PPS is efficient in terms of the Brier score.

The above results continue to hold when at any step t the forecaster has to specify his probability forecasts for a finite number m of events $(E_{t1}, \ldots, E_{tm})$. In this case the prediction is a m-dimensional vector $D_t = (D_{t1}, \ldots, D_{tm})$, and the Brier score is defined as

$$\mathrm{B}_T(D) = \sum_{t=1}^{T}\sum_{j=1}^{m} \{I(E_{tj}) - D_{tj}\}^2 .$$

□

Example 4.4 (Stochastic Linear Regression) Assume that the observations (Y_t) are generated from the following model:

$$Y_t = \theta_1 x_{t1} + \ldots + \theta_p x_{tp} + \epsilon_t, \tag{4.7}$$

where $\theta' = (\theta_1, \ldots, \theta_p)$ is a vector of unknown parameters, and (ϵ_t) is an independent sequence of unobservable errors, each having a Normal distribution with mean zero and variance σ^2. We assume for the moment that σ^2 is known, but later we

will relax this assumption. Also let $x_t' = (x_{t1}, \ldots, x_{tp})$ and $X_t = (x_{jk})_{1 \le j \le t, 1 \le k \le p}$. The regressor vector x_{t+1} may depend on the previous responses and regressors $x_1, y_1, \ldots, x_t, y_t$. Thus, if $\mathcal{F}_t$ is the σ-field generated by X_t and $Y^t = (Y_1, \ldots, Y_t)'$, the vector x_{t+1} is $\mathcal{F}_t$-measurable. Without loss of generality we assume that x_1 is fixed.

Let $\mathbf{B}$ be a Bayesian measure based on a prior distribution for θ which is Normal with mean zero and variance covariance matrix $\lambda^{-1} I$, where I is the identity matrix. Then the predictive distribution under $\mathbf{B}$ for the observation Y_{t+1} is Normal with mean $\tilde{\theta}_t(\lambda)' x_{t+1}$ and variance $\sigma^2 \{1 + x_{t+1}'(\lambda I + X_t' X_t)^{-1} x_{t+1}\}$, where

$$\tilde{\theta}_t(\lambda) = (\lambda I + X_t' X_t)^{-1} X_t Y^t.$$

A PPS based on the predictive means of $\mathbf{B}$ is efficient (see Theorem 4.2) if

$$\sup_t \left\{ x_{t+1}'(\lambda I + X_t' X_t)^{-1} x_{t+1} \right\} < \infty, \ \{\mathbf{P}_\theta, \Theta\}\text{-as.} \tag{4.8}$$

This is a weak condition, since typically the eigenvalues of the matrix $(\lambda I + X_t' X_t)^{-1}$ tend to zero faster than the rate of increase of $\|x_{t+1}\|^2$. If $(X_t' X_t)^{-1}$ exists for $t > t_0$, we can use the fact that the matrix

$$(X_t' X_t)^{-1} - (\lambda I + X_t' X_t)^{-1}$$

is positive definite for every $\lambda > 0$ to show that (4.8) holds if

$$\sup_{t > t_0} \left\{ x_{t+1}'(X_t' X_t)^{-1} x_{t+1} \right\} < \infty, \tag{4.9}$$

or equivalently if

$$\sup_{t > t_0} \left\{ x_t'(X_t' X_t)^{-1} x_t \right\} < 1.$$

Since for any value of λ the PPS is efficient if one of the above conditions holds, then we also have that with probability one for almost all θ

$$\sum_{t=1}^{\infty} \left\{ \tilde{\theta}_t(\lambda_1)' x_{t+1} - \tilde{\theta}_t(\lambda_2)' x_{t+1} \right\}^2 < \infty,$$

for any $\lambda_1, \lambda_2 > 0$. This means that the sequences of predictions from any two ridge estimators are asymptotically equivalent.

The efficiency of the PPS based on the ridge estimators holds for every variance σ^2, and therefore the assumption of a known variance is not necessary, and can be dropped. If the matrix $(X_t'X_t)$ eventually has an inverse, then it can also be shown that a plug-in PPS based on the least squares estimator is efficient and issues predictions asymptotically equivalent to those of any PPS based on a ridge estimator. □

Example 4.5 (Autoregressive Models.) Assume that Y_t follows an autoregressive model of order p, *i.e.*

$$Y_t = \theta_1 Y_{i-1} + \ldots + \theta_p Y_{i-p} + \epsilon_t,$$

where $\{\epsilon_t\}$ are independent Normal with mean 0 and variance σ^2, and $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ is such that all the roots of the characteristic polynomial

$$\phi(z) = z^p - \theta_1 z^{p-1} - \ldots - \theta_p$$

are inside or on the unit circle (non-explosive case). Then we can apply the results of the previous example to show that any PPS based on a ridge estimator or the least squares estimator of θ is efficient. The only difficulty is to show that (4.8) or (4.9) hold with probability one, and this has been established by Lai and Wei (1983).

In the special case of an autoregressive model of order 1, (4.9) holds almost surely regardless of the value of θ (Wei, 1987), and in this case any PPS based on the least squares estimator or any ridge estimator is efficient. In §3.7 it was shown that, for an AR(1) model with $\theta > 1$, any plug-in probability forecasting system is inefficient, and therefore the AR(1) model is an example where an inefficient PFS can produce an efficient PPS. The inefficiency of the plug-in probability forecasting systems is due to their underestimation of the predictive variances. However, this

does not affect the optimality of their implied plug-in point predictions, since all that is required at every step is a good estimate of the conditional mean of the next observation. The fact that these means are linear functions of the parameter θ reduces the problem of optimal prediction to that of optimal estimation. □

Chapter 5

Consistency and Misspecification

5.1 Introduction

In the previous chapters we discussed the issue of prequential efficiency of forecasting systems, as is applied to probability forecasting and point prediction. In this chapter we study another asymptotic property for predictive rules, that of consistency.

Most problems in statistical modelling can be described as follows: a sequence of quantities $Y = (Y_1, Y_2, \ldots)$ is to be observed, and a class of models $\mathcal{M} = \{M(\theta), \theta \in \Theta\}$ indexed by a parameter θ, taking values in a set Θ, is proposed as a suitable description of some properties of Y, which are of interest to the modeller. For example, $\mathcal{M}$ may be a class of probability distributions, or a semi-parametric model that describes the means and variances of (Y_t), or a regression model which models the relationship between different components of Y.

In the majority of the statistical literature, it is assumed that the data (Y_t) are generated by a *data generation process (DGP)* and that there is a value $\theta_0 \in \Theta$ such that $M(\theta_0)$ is the "true" model, in the sense that it describes the properties of interest accurately. This assumption may be reasonable in some situations,

but usually we cannot expect our model to have captured all the properties and relationships among the observed data, which may be very complex. The best we can hope for is that the family of models $\mathcal{M}$ is a good approximation, in some sense which needs to be specified, to the data generating process.

The purpose of any statistical modelling based on the assumption of the existence of a "true" model within our class of models is well defined: we would like to identify the "true" model and make inferences about it. But when $\mathcal{M}$ is a misspecified class of models, then what is the purpose of the modelling, and what are the consequences of misspecification on inferential procedures? It is obvious that in this case we need to think very carefully about the usefulness and limitations of our chosen models.

As a very simple example that clarifies the above point, consider a sequence of i.i.d. observation (Y_t), with mean θ. We believe that the distribution of Y_t is Normal with mean θ, to be estimated from the data, and variance equal to 1. Then, $\mathcal{M} = \{M(\theta) = N(\theta, 1), \theta \in \mathbb{R}\}$, where $N(\cdot, \cdot)$ denotes the Normal distribution. The family $\mathcal{M}$ may not include the true distribution, since it may not be a Normal distribution and also the variance may be different from 1. If we focus our interest on estimating the unknown parameter θ, say by the maximum likelihood estimator $\hat{\theta}_T = (1/T) \sum_{t=1}^{T} Y_t$, then, if the variance is finite, the estimator $\hat{\theta}_T$ is consistent regardless of the true distribution and the value of θ. This consistency property does not mean that we have discovered the "true" distribution of the data, but only that a specific property of this distribution, the mean, can be consistently estimated. Any effort to use the model for other inferential purposes, for example prediction intervals, may be unsuccessful.

It is therefore important when we model some data, using possibly misspecified models, to have a clear understanding of the properties of the data we want to model, to use a statistical methodology that identifies (at least for large samples) the model in $\mathcal{M}$ which is most suitable for our purposes, and to have a good

understanding of the properties of our inferential procedures under misspecification.

In this chapter we adopt the predictivist point of view, which considers a statistical model as a method of making statements about the observable quantities (Y_t). These statements can be phrased as forecasts, and therefore each model can be seen as a predictive system. We will assume therefore that the modeller's aim is to identify the model in $\mathcal{M}$ which issues predictions, which are "closer" to the optimal predictions under the true DGP.

In order to formalize the above thoughts, we need a mathematical framework. Assume that the sequence of variables $Y = (Y_1, Y_2 \ldots)$ are defined on a complete filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbf{P})$. The increasing sequence of σ-subalgebras $(\mathcal{F}_t)$ describes the available information at each time t, and in the simplest case $\mathcal{F}_t = \sigma(Y^t) = \sigma(Y_1, \ldots, Y_t)$. By $E_{t-1}(\cdot)$ we denote the conditional expectation $E(\cdot|\mathcal{F}_{t-1})$.

We try to model the data using a family of models $\mathcal{M} = \{M(\theta), \theta \in \Theta\}$, indexed by the parameter θ taking values in a compact separable metric space (Θ, d). Each model $M(\theta)$ is to be considered as a predictive rule that issues forecasts sequentially, i.e. it may represent a probability forecasting system, or a point prediction system, or any other forecasting model. We let Θ be a metric space in order to allow the results to be applicable in infinite-dimensional parameter spaces. The σ-algebra used to define measurability on Θ is the Borel σ-algebra generated by the open sets of Θ.

Now assume that after we observe the observation Y_t the model $M(\theta)$ is penalized with a $\mathcal{F}_t$-measurable loss function $l_t(\omega, \theta)$, $\omega \in \Omega$, which depends on the forecast of $M(\theta)$, the realized outcome of Y_t, and possibly any other information that is described by $\mathcal{F}_t$. The normalized cumulative loss is denoted $L_T(\omega, \theta) := 1/A_T \sum_{t=1}^{T} l_t(\omega, \theta)$, where (A_T) is a normalizing sequence. We will discuss in the following sections how this sequence should be chosen. The function $L_T(\omega, \theta)$ is a $\mathcal{F}_T$-measurable real function, and represents a statistical criterion, which measures

the empirical predictive performance of the model $M(\theta)$ up to time T.

For a specific set of observed data y^T of Y^T, a standard method of choosing a model from $\mathcal{M}$ is to prefer the one that corresponds to the value of θ that minimizes $L_T(y^T, \theta)$. These estimators are called *extremum* estimators (Gourieroux and Monfort, 1995), and include many different estimators that are currently used in parametric and non-parametric setups, like prediction error estimators, maximum and pseudo-maximum likelihood estimators, least squares estimators etc. Our interest lies in studying the asymptotic behaviour of the extremum estimator $\hat{\theta}_T$ which minimizes $L_T(\omega, \theta)$, i.e. **P**-a.s.

$$L_T(\omega, \hat{\theta}_T) = \min_{\theta \in \Theta} L_T(\omega, \theta).$$

For the study of the performance of these extremum estimators, we need a theory of inference that allows for the possibility of misspecification. Using the fact that in our framework each model is a predictive system, we can replace the notion of a "true" model with that of a "best" model, where by "best" we define the model in $\mathcal{M}$ which issues the best predictions under the true DGP. Then, we would hope that the extremum estimator $\hat{\theta}_T$ would converge, under suitable conditions, to this best value as the number of observations tends to infinity.

The aim of this chapter is to discuss the issue of consistency of extremum estimators for possibly misspecified models. In §5.2 we present a specific theory of inference for misspecified models (White, 1994), which one may apply to our problem, but in §5.3 we show that this theory is not applicable to some non-ergodic models. In §5.4 we show how the theory can be extended using a martingale uniform law of large numbers which we prove in §5.6. In §5.7 we present some examples.

5.2 White's approach

A theory of inference under misspecification has been developed by White and his co-workers (see Gallant and White (1988), White (1994) and references therein).

We give a short description of his approach to the problem of consistency.

The rationale behind White's approach is that since $\hat{\theta}_T$ minimizes $L_T(\omega, \theta)$, then, under the assumption that $L_T(\omega, \theta)$ converges to its overall mean $L_T^*(\theta) := E\{L_T(\omega, \theta)\}$, the estimator $\hat{\theta}_T$ should tend to the value of θ, say θ_T^*, that minimizes $L_T^*(\theta)$. Since Θ indexes a collection of models, and $L_T(\omega, \theta)$ measures the empirical performance of each of these models, then θ_T^* can be interpreted as the value of θ that represents the model that performs best, in this average sense.

In order to prove that the difference between the estimator $\hat{\theta}_T$ and the "best" value θ_T^* tends to zero, as T tends to infinity, the following two assumptions are introduced (Gallant and White, 1988; White, 1994).

Assumption W1. [Identifiable Uniqueness]
For all $\epsilon > 0$,

$$\liminf_{T \to \infty} \left\{ \min_{\theta : d(\theta_T^*, \theta) \geq \epsilon} L_T^*(\theta) - L_T^*(\theta_T^*) \right\} > 0.$$

Assumption W2. [Uniform Law of Large Numbers]
The sequence $L_T(\omega, \theta) - L_T^*(\theta)$ obeys the strong uniform law of large numbers (ULLN):

$$\sup_{\theta \in \Theta} |L_T(\omega, \theta) - L_T^*(\theta)| \to 0 \qquad \mathbf{P}\text{-a.s.}$$

The assumption W1 is used to make sure that the functions $L_T^*(\theta)$ do not become flat around θ_T^*, as T tends to infinity. This assumption can be weakened, as in Davis and Vinter (1985), if we allow the limit of $(\hat{\theta}_T - \theta_T^*)$ to be a set. Using the above assumptions, the following theorem can be established (Gallant and White, 1988; White, 1994; White and Wooldridge, 1991)

Theorem 5.1 *Under the assumptions W1 and W2,* $d(\hat{\theta}_T, \theta_T^*) \to 0$, $\mathbf{P}$*-a.s..*

Although this result can have more general interpretations, in our predictive framework it establishes that the extremum estimator $\hat{\theta}_T$ converges to the value of the

parameter θ which labels the model in the family $\mathcal{M}$ that issues the best predictions in terms of the overall expected predictive loss.

5.3 Counterexamples

The approach described in the previous section fails in cases where the function $L_T(\omega, \theta)$ does not converge to its overall mean, as the following examples show.

Example 5.1 (Stochastic level, Dawid, 1991) Let (X_t), $t \geq 0$, be a sequence of i.i.d. Normal variables with zero mean and unit variance. Let $Y_t = X_0 + X_t$, $t \geq 1$. Our class of models is based on the assumption that $E(Y_t|Y^{t-1}) = \theta$, and therefore according to this model the best prediction for the observation Y_t, in terms of the predictive squared error loss, is θ. If $L_T(\omega, \theta) = (1/T)\sum_{t=1}^{T}(Y_t - \theta)^2$, then $\hat{\theta}_T = (1/T)\sum_{t=1}^{T} Y_t$. The expected loss $L_T^*(\theta)$ is equal to $2 + \theta^2$, and therefore $\theta_T^* = 0$. If White's result were applicable, the estimator $\hat{\theta}_T$ should converge to 0. But, it is easily seen that the estimator $\hat{\theta}_T$ converges almost surely to the observed value x_0 of X_0. In this example the extremum estimator $\hat{\theta}_T$ converges to a data dependent limit. □

Example 5.2 (Mixture of Distributions) A sequence of i.i.d. random variables (Y_t), $t \geq 1$ will be observed, and our parametric family of models $\mathcal{P} = \{\mathbf{P}_\theta, \theta = 1, .., k\}$ consists of a finite number of singular probability distributions for $Y = (Y_1, Y_2, \ldots)$. Let $\mathbf{p}^T(\theta)$ denote the density (with respect to the Lebesgue measure) of the joint distribution for Y^T, and $\mathbf{p}_t(\theta)$ the conditional density of Y_t given Y^{t-1} under $\mathbf{P}_\theta$. Let $L_T(Y^T, \theta) := -(1/T)\log \mathbf{p}^T(\theta) = (1/T)\sum_{t=1}^{T} -\log \mathbf{p}_t(\theta)$. The estimator $\hat{\theta}_T$ is the maximum likelihood estimator of θ.

When the true model belongs to $\mathcal{P}$, then the estimator $\hat{\theta}_T$ converges almost surely to the true value of θ, and is consistent. Assume now that the true model does not lie in $\mathcal{P}$, but it is a mixture of the above models, i.e. $\mathbf{P} = \sum_\theta a_\theta \mathbf{P}_\theta$,

with $\sum_\theta a_\theta = 1$. Let $\mathbf{P}^T$ and $\mathbf{P}_\theta^T$ denote the restrictions of $\mathbf{P}$ and $\mathbf{P}_\theta$ to the first T observations. Then, under $\mathbf{P}$, the expectation of $L_T(Y^T, \theta)$ is minimized at the value θ_T^* which minimizes the Kullback-Leibler distance $K(\mathbf{P}^T, \mathbf{P}_\theta^T)$ (we assume for simplicity that θ_T^* is unique). The sequence of minimizers (θ_T^*) is deterministic, and if White's result were applicable we would expect the estimator $\hat{\theta}_T$ to converge to θ_T^* with probability one under $\mathbf{P}$. But this is not the case, since it is easily seen that $\mathbf{P}\{\hat{\theta}_T \to \theta\} = a_\theta$. This is another case where the estimator $\hat{\theta}_T$ converges to a data dependent limit. □

Example 5.3 (Linear Stochastic Regression) The observed variables (Y_t) are generated from the following model :

$$Y_t = m_t + \epsilon_t$$

where (ϵ_t) is a martingale difference sequence, with respect to an increasing sequence of subalgebras $(\mathcal{F}_t)$, and thus m_t is the conditional mean of Y_t given the past, i.e. $E(\epsilon_t|\mathcal{F}_{t-1}) = 0$, and $m_t = E(Y_t|\mathcal{F}_{t-1})$. Suppose that we try to model Y_t using a linear model

$$E(Y_t|\mathcal{F}_{t-1}) = \theta' x_t,$$

such that $\theta \in \mathbb{R}^p$, and the regressors $x_t \in \mathbb{R}^p$ are $\mathcal{F}_{t-1}$ measurable. When m_t is not equal to $\theta' x_t$, then our model is misspecified. Let $X^T := (x_1, \ldots, x_T)$ and $L_T(Y^T, X^T, \theta) := \sum_{t=1}^T (Y_t - \theta' x_t)^2$. The value of θ that minimizes $L_T(Y^T, X^T, \theta)$ is the least squares estimator $\hat{\theta}_T = (\sum_{t=1}^T x_t x_t')^{-1} \sum_{t=1}^T x_t Y_t$.

In this example the expected value of $L_T(Y^T, X^T, \theta)$ may not exist without further assumptions on the overall expectations of the stochastic regressors (x_t). Even then, the value of θ which minimizes the overall expectation of $L_T(Y^T, X^T, \theta)$ depends on the overall expectations of Y^T and X^T, although the limiting behaviour of $\hat{\theta}_T$ depends on the observed (and not the expected) values of the sequence X^T, as the following lemma shows.

Lemma 5.1 *Let* $\theta_T^{**} = (\sum_{t=1}^T x_t x_t')^{-1} \sum_{t=1}^T x_t m_t$. *Assume that with probability one*

$$\sup_t E(|\epsilon_t|^a | F_{t-1}) < \infty \qquad \textit{(for some } a > 2\textit{)},$$

and that almost surely $\lambda \min(t) \to \infty$, *and* $\log\{\lambda \max t\} = o\{\lambda \min(t)\}$, *where* $\lambda \min(t)$ *and* $\lambda \max t$ *are the minimum and maximum eigenvalues of* $\sum_{t=1}^T x_t x_t'$. *Then with probability one under* $\mathbf{P}$ $\|\hat{\theta}_T - \theta_T^{**}\| \to 0$.

Proof. The estimator $\hat{\theta}_T$ is equal to:

$$\begin{aligned} \hat{\theta}_T &= (\sum_{t=1}^T x_t x_t')^{-1} \sum_{t=1}^T x_t y_t \\ &= (\sum_{t=1}^T x_t x_t')^{-1} \sum_{t=1}^T x_t (m_t + \epsilon_t) \\ &= \theta_T^{**} + (\sum_{t=1}^T x_t x_t')^{-1} \sum_{t=1}^T x_t \epsilon_t. \end{aligned}$$

Now using the results in Lai and Wei (1982), it can be shown that the difference $\|\hat{\theta}_T - \theta_T^{**}\|$ converges to zero. □

In all the above examples, we see that under misspecification the estimator $\hat{\theta}_T$ converges to a stochastic limit. Two questions arise in this case. First, what is the interpretation of such a limit, and second, how can we extend the theory in order to cover these cases as well? In the next section we propose some answers to these questions.

5.4 An alternative view of consistency

As we discussed in the introduction of this chapter, for each one of the examples in the previous section the loss function $L_T(\omega, \theta)$ is interpreted as a cumulative measure of the predictive ability of a statistical model. White's result suggests that we should expect $\hat{\theta}_T$ to converge to the value of θ that minimizes the predictive risk $L_T^*(\theta)$, which is based on an overall expectation. This expected loss does

not take into account the sequential nature of prediction, and the fact that the optimal predictor for the observation Y_t, under the true distribution $\mathbf{P}$, depends on the observed information up to time $t-1$. Under misspecification for different sequences of observed data we may have different one step ahead optimal predictions, and therefore the best approximation (in terms of predictive ability) to the data generating process, from within our family of models, may be dependent on the observed sequence of data. This will be the case for non-ergodic models, and for observations with long-term dependencies that do not die sufficiently fast.

An alternative view, is to try to relate the behaviour of the loss function $L_T(\omega,\theta)$ with that of the sum of the conditionally expected one-step ahead prediction losses (Dawid, 1991). The same idea is also briefly discussed in Caines (1988), but is not explored further there.

In mathematical terms since $L_T(\omega,\theta) = (1/A_T)\sum_{t=1}^{T} l_t(\omega,\theta)$, for some $\mathcal{F}_t$-measurable functions $l_t(\omega,\theta)$, then it may be more relevant to try to compare $\hat{\theta}_T$ with the sequence (θ_T^{**}) of minimizers of the function

$$L_T^{**}(\omega,\theta) := (1/A_T)\sum_{t=1}^{T} E_{t-1}\{l_t(\omega,\theta)\},$$

instead of

$$L_T^{*}(\theta) = E\{(1/A_T)\sum_{t=1}^{T} l_t(\omega,\theta)\},$$

since the conditional expectation $E_{t-1}\{l_t(\omega,\theta)\}$ is the conditional predictive risk, based on all observations up to time $t-1$, whereas the overall expectation $E\{l_t(\omega,\theta)\}$ is the unconditional predictive risk which does not take into account the observed data up to time $t-1$. This approach may also allow us to use a non-deterministic sequence A_T for the denominator. This can be very useful since usually the sequence A_T is related to the information available in the data, and for some models the growth of this information is stochastic, and varies for different sequences.

Using this approach we can re-examine examples 5.1-5.3, and give a natural interpretation to the limiting behaviour of $\hat{\theta}_T$.

Example 5.1 (contd.) Let $\mathcal{F}_T$ be the algebra generated by $(Y_1, \ldots, Y_T)$. Observe that although the overall predictive risk $E\{(1/T)\sum_{t=1}^{T}(Y_t - \theta)^2\}$ is minimized for $\theta = 0$, the average one-step ahead predictive risk $(1/T)\sum_{t=1}^{T} E_{t-1}\{(Y_t - \theta)^2\}$ is minimized for $\theta_T^{**} = (1 - 1/T)\hat{\theta}_T$, and $|\hat{\theta}_T - \theta_T^{**}| \to 0$. □

Example 5.2 (contd.) Although under each $\mathbf{P}_\theta$, the observations (Y_t) are independent identically distributed, under the mixture $\mathbf{P}$, the variables (Y_t) are exchangeable, but not independent. Let $\mathcal{F}_t = \sigma(Y_1, Y_2, \ldots, Y_t)$, $l_t(\omega, \theta) = -\log \mathbf{p}_t(\theta)$ and observe that $L_T(\theta) = (1/T)\sum_{t=1}^{T} l_t(\omega, \theta)$. Denote by $\mathbf{p}_{t+1}$ the conditional distribution of Y_{t+1} given Y^t under $\mathbf{P}$. Then, it is well known that $\mathbf{p}_{t+1} = \sum_\theta a_t(\theta)\mathbf{p}_{t+1}(\theta)$, where $\{a_t(\theta)\}$ is the posterior distribution of θ given the observations Y^t. Since the distributions $\mathbf{P}_\theta$ are singular, then $\mathbf{P}$-a.s. the posterior probability $a_t(\hat{\theta}_t)$ converges to one. This implies that the Kullback distance $K(\mathbf{p}_{t+1}, \mathbf{p}_{t+1}(\hat{\theta}_t))$ converges to 0, and for all $\theta \neq \lim_t \hat{\theta}_t$. it stays positive, i.e. $\liminf_t K(\mathbf{p}_{t+1}, \mathbf{p}_{t+1}(\theta)) > 0$.

The function $L_T^{**}(\omega, \theta) = (1/T)\sum_{t=1}^{T} E_{t-1}\{l_t(\omega, \theta)\}$ is minimized at the same value that minimizes the function $(1/T)\sum_{t=1}^{T} K(\mathbf{p}_t, \mathbf{p}_t(\theta))$, which, as was discussed above, is asymptotically minimized at the value $\lim_T \hat{\theta}_T$. Therefore $|\hat{\theta}_T - \theta_T^{**}| \to 0$. □

Example 5.3 (contd). Observe that $\theta_T^{**} = (\sum_{t=1}^{T} x_t x_t')^{-1} \sum_{t=1}^{T} x_t m_t$ is the value of θ that minimizes $L_T^{**}(Y^T, X^T, \theta) = \sum_{t=1}^{T} E_{t-1}\{(Y_t - \theta' x_t)^2\}$. □

In each one of the examples the extremum estimator $\hat{\theta}_T$ converges to the value of θ that minimizes the sum of the conditional one step ahead predictive risks. We will present an extension of White's theory that can cover these cases.

5.5 A General Consistency Theorem

Based on the ideas of the previous section, we can now present a general theorem on the behaviour of extremum estimators under model misspecification.

First we need to establish the existence and measurability of the estimators $\hat{\theta}_T$ and θ_T^{**}.

5.5.1 Existence

We introduce the following assumptions:

Assumption E1. The metric space (Θ, d) is compact and separable.

Assumption E2. (a) For every t, the loss function $l_t(\cdot, \theta) : \Omega \to \mathbb{R} \cup \{-\infty, \infty\}$ is $\mathcal{F}_t$-measurable, for each θ in Θ. The function $l_t(\omega, \cdot)$ is continuous on Θ almost surely, i.e. it is continuous for all ω in an event $F_t \in \mathcal{F}_t$ such that $\mathbf{P}(F_t) = 1$.

(b) For every t, the function $E_{t-1}\{l_t(\cdot, \theta)\} : \Omega \to \mathbb{R} \cup \{-\infty, \infty\}$ is $\mathcal{F}_{t-1}$-measurable, for each θ in Θ. The function $E_{t-1}\{l_t(\omega, \cdot)\}$ is continuous on Θ almost surely, i.e. it is continuous for all ω in an event $F_{t-1} \in \mathcal{F}_{t-1}$ such that $\mathbf{P}(F_{t-1}) = 1$.

When the assumptions $E1$ and $E2$ hold, then almost surely the estimators $\hat{\theta}_T$ and θ_T^{**} exist, and are measurable as the following lemma shows (Gallant and White, 1988; White, 1994; White and Wooldridge, 1991).

Lemma 5.2 *Let $(\Omega, \mathcal{F})$ be a measurable space, and let (Θ, d) be a compact, separable metric space. Let $Q : \Omega \times \Theta \to \mathbb{R} \cup \{-\infty, \infty\}$ be such that $Q(\cdot, \theta)$ is $\mathcal{F}$-measurable for each θ in Θ, and $Q(\omega, \cdot)$ is continuous for all ω in an event $F \in \mathcal{F}$. Then there exists a function $\hat{\theta} : \Omega \to \Theta$ such that $\hat{\theta}$ is $\mathcal{F}$-measurable and for all ω in F*

$$Q(\omega, \hat{\theta}(\omega)) = \inf_{\theta \in \Theta} Q(\omega, \theta).$$

5.5.2 Consistency

In order to prove that $\hat{\theta}_T$ converges to θ_T^{**} we need the following assumptions, which are modified versions of conditions W1 and W2.

Assumption C1. [Asymptotic Identifiability]

The function $L_T^{**} : \Omega \times \Theta \to \mathbb{R} \cup \{-\infty, \infty\}$ is $\mathcal{F}_{T-1}$ measurable, and **P**-a.s. has a minimum on Θ at $\theta_T^{**}(\omega)$, for every T sufficiently large. Let $\epsilon > 0$ and $B_T^c(\epsilon) := \{\theta \in \Theta : d(\theta, \theta_T^{**}) \geq \epsilon\}$. Then **P**-a.s.

$$\liminf_{T\to\infty} \Big\{ \min_{\theta \in B_T^c(\epsilon)} L_T^{**}(\omega, \theta) - L_T^{**}(\omega, \theta_T^{**}) \Big\} > 0. \tag{5.1}$$

Assumption C2. [Martingale Uniform Law of Large Numbers]
With probability one under **P**,

$$\sup_{\theta \in \Theta} |L_T(\omega, \theta) - L_T^{**}(\omega, \theta)| \to 0. \tag{5.2}$$

Theorem 5.2 *Assume that conditions C1, C2 hold. Let $\hat{\theta}_T$ be an estimator that,* **P***-a.s., minimizes $L_T(\omega, \theta)$, for all T sufficiently large. Then, with probability one under* **P**,

$$d(\hat{\theta}_T, \theta_T^{**}) \to 0.$$

Proof of theorem 5.2. The following events have all probability one under **P**:

$$\begin{aligned}
F_1 &= \{\omega \in \Omega : L_T^{**}(\omega, \theta) \text{ has a unique minimum at } \theta_T^{**} \text{ for all } T \text{ sufficiently large}\},\\
F_2 &= \{\omega \in \Omega : L_T(\omega, \theta) \text{ has a unique minimum at } \hat{\theta}_T \text{ for all } T \text{ sufficiently large}\},\\
F_3 &= \{\omega \in \Omega : \text{ for all } \epsilon > 0, \liminf_{T\to\infty} \left(\min_{\theta \in B_T^c(\epsilon)} L_T^{**}(\omega, \theta) - L_T^{**}(\omega, \theta_T^{**}) \right) > 0\},\\
F_4 &= \{\omega \in \Omega : \sup_{\theta \in \Theta} |L_T(\omega, \theta) - L_T^{**}(\omega, \theta)| \to 0\}.
\end{aligned}$$

It follows that the event $F := F_1 \bigcap F_2 \bigcap F_3 \bigcap F_4$ has also probability one under **P**. Given $\epsilon > 0$, for all ω in F, there is $T_1 := T_1(\omega, \epsilon) < \infty$, such that

$$\delta(\epsilon) := \inf_{T > T_1} \left(\min_{\theta \in B_T^c(\epsilon)} L_T^{**}(\omega, \theta) - L_T^{**}(\omega, \theta_T^{**}) \right) > 0.$$

Also, for all $\omega \in F$, and all $T > T_2(\omega, \delta(\epsilon))$

$$|L_T(\omega, \theta_T^{**}) - L_T^{**}(\omega, \theta_T^{**})| < \delta(\epsilon)/2,$$

so that

$$L_T^{**}(\omega, \theta_T^{**}) > L_T(\omega, \theta_T^{**}) - \delta(\epsilon)/2 \geq L_T(\omega, \hat{\theta}_T) - \delta(\epsilon)/2,$$

and for all $\omega \in F$ and $T > T_3(\omega, \delta(\epsilon))$

$$|L_T^{**}(\omega, \hat{\theta}_T) - L_T(\omega, \hat{\theta}_T)| < \delta(\epsilon)/2.$$

Then

$$L_T^{**}(\omega, \hat{\theta}_T) - L_T^{**}(\omega, \theta_T^{**}) \leq L_T^{**}(\omega, \hat{\theta}_T) - L_T(\omega, \hat{\theta}_T) + \delta(\epsilon)/2 < +\delta(\epsilon)/2 + \delta(\epsilon)/2 = \delta(\epsilon),$$

and it follows that $d(\theta_T^{**}, \hat{\theta}_T) < \epsilon$, for all $\omega \in F$, and all $T > \max\{T_1, T_2, T_3\}$. Since ϵ is arbitrary, and $\mathbf{P}(F) = 1$, it follows that $\mathbf{P}$-a.s.

$$d(\hat{\theta}_T, \theta_T^{**}) \to 0.$$

□

5.6 A Uniform Law of Large Numbers for Martingales

The main difference between our approach and White's approach is that we replace the uniform law of large numbers with a martingale uniform law of large numbers. To the best of our knowledge such a law has not been proven yet, and our aim in this section is to present sufficient conditions for a martingale ULLN, which can be used to verify condition C2 in order to establish consistency.

Our approach is based on a modification of the generic laws of large numbers presented by Andrews(1987, 1992), which can not be applied directly for reasons that will become apparent later.

To give a more general result assume that $(Y_t, t \geq 1)$ is a sequence of stochastic elements, taking values in a set $\mathcal{Y}$, defined on a complete filtered probability space $\{\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbf{P}\}$. For each t, $l_t(Y_t, \theta)$ is a measurable function from $\mathcal{Y} \times \Theta$ to $\mathbb{R}$, for some metric space (Θ, d). Let $B(\theta, \rho)$ be the open ball around θ of radius ρ. Define:

$$\overline{l_t}(Y_t, \theta, \rho) = \sup_{s \in B(\theta, \rho)} l_t(Y_t, s),$$

$$\underline{l}_t(Y_t, \theta, \rho) = \inf_{s \in B(\theta,\rho)} l_t(Y_t, s).$$

Let A_T be a predictable increasing sequence with $A_1 \geq 1$ and $\lim_T A_T = \infty$. A sequence of real random variables (Z_t) is said to satisfy a pointwise martingale strong law of large numbers (with denominator A_T) if, with probability one,

$$\lim_{T\to\infty} \frac{1}{A_T} \sum_{t=1}^{T} \{Z_t - E_{t-1}(Z_t)\} = 0.$$

One of the main reasons why Andrews's result is not applicable to our case is that we want to use a predictable, and not a deterministic, denominator.

In order to prove the main result we introduce the following assumptions.

Assumption U1. The metric space (Θ, d) is compact.

Assumption U2. For any $\theta \in \Theta$ there is $\rho(\theta)$ such that for all $\rho < \rho(\theta)$ the sequence of random variables $(\overline{l}_t(Y_t, \theta, \rho))$ and $(\underline{l}_t(Y_t, \theta, \rho))$ satisfy pointwise strong martingale LLN's (with common denominator an increasing predictable sequence A_T).

Assumption U3. For all $\theta \in \Theta$, **P**-a.s.,

$$\lim_{\rho\to 0} \limsup_{T\geq 1} \frac{1}{A_T} \sum_{t=1}^{T} E_{t-1}\{\overline{l}_t(Y_t, \theta, \rho) - \underline{l}_t(Y_t, \theta, \rho)\} = 0.$$

Although our assumptions are similar to Andrews's (1987) assumptions, they are weaker because the use of conditional expectations, instead of unconditional expectations, weakens the degree of dependence that the conditions impose on the variables (Y_t). Using the above conditions we can now prove the following theorem:

Theorem 5.3 (Martingale ULLN) *If Assumptions U1-U3 hold, then* **P**-*a.s.,*

$$\sup_{\theta\in\Theta} \left| \frac{1}{A_T} \sum_{t=1}^{T} [l_t(Y_t, \theta) - E_{t-1}\{l_t(Y_t, \theta)\}] \right| \to 0.$$

Proof of theorem 5.3. The proof will follow the same method as in Andrews (1987). Using assumption U3, for a given $\epsilon > 0$ and $\theta \in \Theta$, there is an event $F(\theta)$,

with $P\{F(\theta)\} = 1$, such that for all $\omega \in F(\theta)$ we can choose $\rho = \rho(\theta) > 0$ such that for all $T \geq T_1(\omega, \theta)$,

$$\frac{1}{A_T}\sum_{t=1}^{T} E_{t-1}\{\,\overline{l}_t(Y_t, \theta, \rho) - \underline{l}_t(Y_t, \theta, \rho)\,\} < \epsilon.$$

The collection of balls $\{B(\theta, \rho(\theta)), \theta \in \Theta\}$, is an open cover of the compact set Θ, and therefore has a finite subcover $\{B(\theta_j, \rho(\theta_j)) : j = 1, 2, .., J\}$.

Let $F_0 = \bigcap_{j=1}^{J} F(\theta_j)$. For all $\omega \in F_0$, and any $s \in B(\theta_1, \rho(\theta_1))$, we have for all $T > max_j T_1(\omega, \theta_j)$

$$\begin{aligned}
\frac{1}{A_T}\sum_{t=1}^{T}[\,l_t(Y_t, s) - E_{t-1}\{\,l_t(Y_t, s)\,\}] &\leq \frac{1}{A_T}\sum_{t=1}^{T}\Big(\overline{l}_t(Y_t, \theta_1, \rho(\theta_1)) - E_{t-1}[\,\underline{l}_t(Y_t, \theta_1, \rho(\theta_1))\,]\Big) \\
&\leq \frac{1}{A_T}\sum_{t=1}^{T}\Big(\overline{l}_t(Y_t, \theta_1, \rho(\theta_1)) - E_{t-1}[\,\overline{l}_t(Y_t, \theta_1, \rho(\theta_1))\,]\Big) + \epsilon
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{A_T}\sum_{t=1}^{T}[\,l_t(Y_t, s) - E_{t-1}\{\,l_t(Y_t, s)\,\}] &\geq \frac{1}{A_T}\sum_{t=1}^{T}\Big(\underline{l}_t(Y_t, \theta_1, \rho(\theta_1)) - E_{t-1}[\,\overline{l}_t(Y_t, \theta_1, \rho(\theta_1))\,]\Big) \\
&\geq \frac{1}{A_T}\sum_{t=1}^{T}\Big(\underline{l}_t(Y_t, \theta_1, \rho(\theta_1)) - E_{t-1}[\,\underline{l}_t(Y_t, \theta_1, \rho(\theta_1))\,]\Big) - \epsilon
\end{aligned}$$

Then for every $\omega \in F_0$ and $\theta \in \Theta$,

$$\begin{aligned}
&\min_{j \leq J}\frac{1}{A_T}\sum_{t=1}^{T}\Big(\underline{l}_t(Y_t, \theta_j, \rho(\theta_j)) - E_{t-1}[\,\underline{l}_t(Y_t, \theta_j, \rho(\theta_j))\,]\Big) - \epsilon \\
\leq\ &\frac{1}{A_T}\sum_{t=1}^{T}[\,l_t(Y_t, \theta) - E_{t-1}\{\,l_t(Y_t, \theta)\,\}] \\
\leq\ &\max_{j \leq J}\frac{1}{A_T}\sum_{t=1}^{T}\Big(\overline{l}_t(Y_t, \theta_j, \rho(\theta_j)) - E_{t-1}[\,\overline{l}_t(Y_t, \theta_j, \rho(\theta_j))\,]\Big) + \epsilon.
\end{aligned}$$

From assumption U2 the above upper and the lower limits converge to ϵ and $-\epsilon$ respectively for all ω in an event F_1 which has probability one. Since $\epsilon > 0$ is arbitrary, and $P(F_0 \bigcap F_1) = 1$ the proof of the theorem is complete. □

In most cases assumptions U1-U3 are difficult to verify, and then we may use the following assumptions:

Assumption U4. For each $\theta \in \Theta$, there is a constant $\tau > 0$ such that $d(\theta, s) \leq \tau$ implies that for every $t \geq 1$, **P**-a.s.,

$$| l_t(Y_t, \theta) - l_t(Y_t, s) | \leq B_t(Y_t)\, h\{d(\theta, s)\},$$

where $\{B_t(\cdot)\}$ is a sequence of $\mathcal{F}_t$-measurable functions such that **P**-a.s.,

$$\limsup_T \frac{1}{A_T} \sum_{t=1}^{T} E_{t-1}\{B_t(Y_t)\} < \infty,$$

and $h(\cdot)$ is a non-random function such that $h(y) \downarrow h(0) = 0$ as $y \downarrow 0$. The null sets, and also $t, B_t(\cdot)$, and h may depend on θ.

Assumption U5. Θ is subset of $\mathbb{R}^p$, $l_t(Y_t, \theta)$ is differentiable with respect to θ in a neighborhood of θ_0, **P**-a.s., for every t and for all $\theta_0 \in \Theta^*$, where Θ^* is some convex open set that contains Θ. Also $\partial\, l_t(Y_t, \theta)\,/\partial\theta$ and $sup_{t\in\Theta^*} \| \partial\, l_t(Y_t, s)\,/\partial\theta \|$ are random variables for any $\theta \in \Theta$, and $t \geq 1$, and **P**-a.s.,

$$\limsup_{T-\infty} \frac{1}{A_T} \sum_{t=1}^{T} E_{t-1}\Big\{ sup_{t\in\Theta^*} \| \partial\, l_t(Y_t, s)\,/\partial\theta \| \Big\} < \infty.$$

Assumption U6. With **P** probability one

$$\sum_{t=1}^{\infty} \frac{E_{t-1}\{\sup_{\theta\in\Theta} | l_t(Y_t, \theta) |^2\}}{A_t} < \infty.$$

Assumption U7. There is $\epsilon > 0$ such that **P**-a.s.

$$\Big(\sum_{t=1}^{T} E_{t-1}\{\sup_{\theta\in\Theta} | l_t(Y_t, \theta) |^2\} \Big)^{(1+\epsilon)/2} = O(A_T).$$

Lemma 5.3 *The following hold :*

(*a*) *Assumption U4 implies Assumption U3.*

(*b*) *Assumption U5 implies Assumption U4.*

(*c*) *Assumption U6 implies Assumption U2.*

(*d*) *Assumption U7 implies Assumption U2.*

Proof of lemma 5.3.

(*a*) **Assumption U4 implies Assumption U3.**

Let $\theta \in \Theta$. Then we can see that U4 implies U3 as,

$$\limsup_{\rho \to 0} \sup_{n \geq 1} \frac{1}{A_T} \sum_{t=1}^{T} E_{t-1}\{ \overline{l}_t(Y_t, \theta, \rho) - \underline{l}_t(Y_t, \theta, \rho) \}$$

$$\leq \limsup_{\rho \to 0} \sup_{T \geq 1} \frac{1}{A_T} \sum_{t=1}^{T} E_{t-1}\{ | \overline{l}_t(Y_t, \theta, \rho) - l_t(Y_t, \theta) | + | l_t(Y_t, \theta) - \underline{l}_t(Y_t, \theta, \rho) | \}$$

$$\leq 2 \lim_{\rho \to 0} h(\rho) \sup_{T \geq 1} \frac{1}{A_T} \sum_{t=1}^{T} E_{t-1}\{B_t(Y_t)\} = 0.$$

□

(*b*) **Assumption U5 implies Assumption U4.**

In order to establish that U5 implies U4, we can use the mean value theorem to show that, **P**-a.s.,

$$| l_t(Y_t, s) - l_t(Y_t, \theta) | \leq sup_{\theta^* \in \Theta^*} \| \partial l_t(Y_t, \theta^*)/\partial \theta \| \cdot \| s - \theta \|$$

Then by setting $h(y) = y$ and

$$B_t(Y_t) = sup_{\theta^* \in \Theta^*} \| \partial l_t(Y_t, \theta^*)/\partial \theta \|$$

we get U4. □

(*c*) **Assumption U6 implies Assumption U2.**

For every θ and ρ small enough we have,

$$\sup_{\theta \in \Theta} | l_t(Y_t, \theta) |^2 \geq | \overline{l}_t(Y_t, \theta, \rho) |^2$$

and therefore

$$\frac{E_{t-1}\{\sup_{\theta \in \Theta} | l_t(Y_t, \theta) |^2\}}{A_t^2} \geq \frac{E_{t-1}\{| \overline{l}_t(Y_t, \theta, \rho) |^2\}}{A_t^2}.$$

Thus,

$$\sum_{i=1}^{n} \frac{V_{t-1}\{ \overline{l}_t(Y_t, \theta, \rho) \}}{A_t^2} < \infty,$$

where $V_{t-1}(\cdot)$ is the conditional variance given $\mathcal{F}_{t-1}$. Using the martingale SLLN for square integrable martingales (Shiryayev, 1996) we get that

$$\frac{1}{A_T}\sum_{t=1}^{T}[\,\overline{l}_t(Y_t,\theta,\rho) - E_{t-1}\{\,\overline{l}_t(Y_t,\theta,\rho)\,\}] \to 0.$$

The same argument can be used for $\underline{l}_t(Y_t,\theta,\rho)$ and $l_t(Y_t,\theta)$. □

(d) **Assumption U7 implies Assumption U2.**

We will use a similar method as in *(c)*. For every θ and ρ small enough we have,

$$\sup_{\theta\in\Theta} |\, l_t(Y_t,\theta)\,|^2 \geq |\,\overline{l}_t(Y_t,\theta,\rho)\,|^2$$

and therefore

$$\sum_{t=1}^{T} E_{t-1}\{\sup_{\theta\in\Theta} |\, l_t(Y_t,\theta)\,|^2\} \geq \sum_{t=1}^{T} V_{t-1}\{\,\overline{l}_t(Y_t,\theta,\rho)\,\}, \tag{5.3}$$

as in *(c)* above. From (Lai and Wei, 1982) we know that

$$\sum_{t=1}^{T}[\,\overline{l}_t(Y_t,\theta,\rho) - E_{t-1}\{\,\overline{l}_t(Y_t,\theta,\rho)\,\}] = o\Big(\Big\{\sum_{t=1}^{T} V_{t-1}\{\,\overline{l}_t(Y_t,\theta,\rho)\,\}\Big\}^{(1+\epsilon)/2}\Big).$$

Using (5.3) and the fact that

$$\Big(\sum_{t=1}^{T} E_{t-1}\{\sup_{\theta\in\Theta} |\, l_t(Y_t,\theta)\,|^2\}\Big)^{(1+\epsilon)/2} = O(A_T).$$

we get

$$\frac{1}{A_T}\sum_{t=1}^{T}[\,\overline{l}_t(Y_t,\theta,\rho) - E_{t-1}\{\,\overline{l}_t(Y_t,\theta,\rho)\,\}] \to 0.$$

The same argument can be used for $\underline{l}_t(Y_t,\theta,\rho)$ and $l_t(Y_t,\theta)$. □

5.7 Examples

In this section we present some examples, where the results of this chapter can be applied, in order to highlight some points on consistency for misspecified models.

Example 5.4 (AR(1) model) Let $Y_0 = 0$ and assume that

$$Y_t = \theta_0 Y_{t-1} + \epsilon_t,$$

where (ϵ_t) is a martingale difference sequence with constant variance. We fit the model :

$$E(Y_t|Y^{t-1}) = \theta,$$

and estimate θ using least squares. Then, it is easily seen that $\hat{\theta}_T = \bar{Y}_T$, $\theta_T^* = 0$, and $\theta_T^{**} = \theta_0(1 - 1/T)\bar{Y}_{T-1}$. If $|\theta_0| < 1$, then the true model is stationary, and $\lim_T \hat{\theta}_T = \lim_T \theta_T^* = \lim_T \theta_T^{**} = 0$. When $|\theta_0| \geq 1$, the true model is not stationary, and $\bar{Y}_T$ does not converge to θ_T^*. But, $\hat{\theta}_T - \theta_T^{**} = (1/T)\sum_{t=1}^T \epsilon_t$, and therefore $|\hat{\theta}_T - \theta_T^{**}| \to 0$. This example shows that our approach can be applied to non-stationary, non-ergodic models. □

The next example shows that under misspecification different loss functions lead to different estimators, and therefore we should be careful to choose the appropiate loss function for our prediction/decision problem.

Example 5.5 (AR(2) model) Let $Y_0 = Y_{-1} = 0$ and assume that (Y_t) follows the following AR(2) stationary model:

$$Y_t = \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \epsilon_t,$$

where (ϵ_t) is a martingale difference sequence with constant variance. Assume that we use the model:

$$E(Y_t|Y^{t-1}) = \theta Y_{t-1},$$

and, after we observe the data Y^T, we are interested in predicting two steps ahead in the future.

There are two methods of obtaining a two-step ahead prediction. One is to estimate θ using least squares, and then substitute θ in the formula $E(Y_{T+2}|Y^T) =$

$\theta^2 Y_T$ with the least squares estimator, or we can estimate $\phi = \theta^2$ directly by minimizing the 2-step ahead prediction errors $\sum_{t=1}^{T}(Y_t - \phi Y_{t-2})^2$.

If we use Lemma 5.1, we see that by minimizing the one step ahead prediction errors $\sum_{t=1}^{T}(Y_t - \theta Y_{t-1})^2$, our estimate $\hat{\theta}_{1,T}$ converges to $\rho(1)$, i.e. the autocorrelation at lag one. Then, for large T, our 2-step ahead prediction is approximately $\hat{Y}_{1,T+2} = \rho(1)^2 Y_T$. Using the same lemma, we see that if we use the second method with an unrestricted value for ϕ, then $\hat{\phi}$ converges to the autocorrelation at lag 2, $\rho(2)$, and asymptotically our prediction is $\hat{Y}_{2,T+2} = \rho(2)Y_T$. Since $\rho(2)$ may be different from $\rho(1)^2$, we see that different loss functions result in different estimators of θ, and therefore different predictions. It is also clear that the second method gives the best predictions.

The explanation for the above result is simple. The definition of what is the "best" value of θ under misspecification depends on the loss function we use. If we minimize the one step ahead prediction errors, then our estimator converges to the value of θ that issues the best one step ahead predictions. If, on the other hand, we minimize the two steps ahead prediction errors then the estimator converges to the value of θ that issues the best two step ahead predictions. Although in a well specified model the two values are the same, i.e. the true value of θ, in a misspecified model different loss functions give different approximations to the true model. We should be careful therefore first to specify the decision problem we want to solve, and then to estimate θ. □

Example 5.6 (Error in variables) Assume that the variables (Y_t) are generated from the model:

$$Y_t = \theta_0 x_t + \epsilon_t,$$

where (ϵ_t) is a sequence of i.i.d. variables with finite variance, and (x_t) is a sequence of random variables such that $(1/T)\sum_{t=1}^{T} x_t^2$ converges to a positive random variable X. Instead of the sequence (x_t) we observe the sequence (z_t), such that $z_t = x_t + v_t$, where (v_t) is a sequence of variables, independent of (x_t), with mean zero and such

that $(1/T)\sum_{t=1}^{T} v_t^2$ converges to a positive random variable V. We may think of v_t as the error in measuring the regressor x_t. If we model the data using the model

$$E(Y_t|z_t) = \theta z_t,$$

and estimate θ using least squares, then we get $\hat{\theta}_T = \sum_{t=1}^{T}(y_t z_t)/\sum_{t=1}^{T} z_t^2$, $\theta_T^{**} = \theta_0 \sum_{t=1}^{T}(x_t z_t)/\sum_{t=1}^{T} z_t^2$, and using Lemma (5.1) we have $|\hat{\theta}_T - \theta_T^{**}| \to 0$. Since

$$\theta_T^{**} = \theta_0 \frac{\sum_{t=1}^{T}(x_t z_t)}{\sum_{t=1}^{T} z_t^2} = \theta_0 \frac{\sum_{t=1}^{T}(x_t^2 + x_t v_t)}{\sum_{t=1}^{T}(x_t^2 + v_t^2 + 2x_t v_t)} \to \theta_0 \frac{X}{X+V},$$

we see that $\hat{\theta}_T$ is estimation inconsistent, in the sense that it does not converges to the value θ_0, but it is prediction consistent since it converges to the value of θ that issues asymptotically the best one step ahead predictions. □

The next example presents sufficient conditions for the consistency of least squares estimators in non-linear stochastic regression models, when the model is well specified. We present this example to show that our approach is useful in cases where a model is specified using a martingale structure, and also because it illustrates how the techniques of this chapter can be adopted to specific problems.

Example 5.7 (Non-Linear Stochastic Regression Models) Assume that the data (Y_t) are defined on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbf{P})$, and are generated from the non-linear stochastic regression model:

$$Y_t = f_t(\theta) + \epsilon_t, \tag{5.4}$$

where for every t, $f_t(\theta)$ is a $\mathcal{F}_{t-1}$-measurable function of $\theta \in \mathbb{R}^p$, the parameter θ takes values in a compact parameter set $\Theta \subset \mathbb{R}^p$, and the sequence of errors $\{\epsilon_t\}$ is a martingale difference sequence with respect to $(\mathcal{F}_t)$ such that $\mathbf{P}$-a.s.,

$$\sup_t E_{t-1}(\epsilon_t^2) < \infty. \tag{5.5}$$

The above class of models is very general, and for example $f_t(\theta)$ can be a function of the past observations and other exogenous inputs. Many nonlinear models,

such as nonlinear autoregressive models with exogenous regressors, used in time series, adaptive control, stochastic approximation, and sequential design, obey model (5.4).

The unknown parameter θ can be estimated using the least squares estimator which is defined as the parameter value $\hat{\theta}_T$ that minimizes the sum of squared errors

$$S_T(\theta) := \sum_{t=1}^{T} \{y_t - f_t(\theta)\}^2.$$

The strong consistency of the estimate $\hat{\theta}_T$ is a very important problem, especially for identification and control. This property has been studied extensively for linear and nonlinear regreesion models (Anderson and Taylor, 1979; Christopeit and Helmes, 1980; Wu, 1981; Lai and Wei, 1982; Lai, 1994), but consistency for nonlinear stochastic regression models has been proven under strict conditions by Lai (1994), which include smoothness conditions on all partial derivatives up to order p. In this example we use an approach based on Theorem 5.2 which does not make any assumptions on the existence of derivatives. We introduce the following assumptions:

Assumption SR1 The parameter set Θ is a compact subset of $\mathbb{R}^p$.

Assumption SR2 Let θ_0 denote the true value of θ. For every $\lambda \neq \theta_0$ there exists $1 < \rho_\lambda < 2$, and an open ball centered at λ (denoted by $B(\lambda)$) such that:

$$A_T := \inf_{s \in B(\lambda)} \sum_{t=1}^{T} \{f_t(s) - f_t(\theta_0)\}^2 \to \infty \qquad \mathbf{P}\text{-a.s.}, \tag{5.6}$$

$$\sum_{t=1}^{T} \sup_{s \in B(\lambda)} \{f_t(s) - f_t(\theta_0)\}^2 = O\left(A_T^{\rho_\lambda}\right) \qquad \mathbf{P}\text{-a.s.}, \tag{5.7}$$

and also there is a sequence of $\mathcal{F}_t$-measurable variables $M_t(\lambda)$ such that for all s_1, s_2 in $B(\lambda)$,

$$|f_t(s_1) - f_t(s_2)| \leq h(\|s_1 - s_2\|)\, M_t(\lambda), \tag{5.8}$$

and

$$\sum_{t=1}^{T} E_{t-1}\{M_t(\lambda)\} = O(A_T) \qquad \mathbf{P}\text{-a.s.}, \tag{5.9}$$

where $h(\cdot)$ is a non-random function such that $h(y) \downarrow h(0) = 0$, as $y \downarrow 0$.

Theorem 5.4 *Let the assumptions SR1 and SR2 hold. Then with probability one* $\|\hat{\theta}_T - \theta_0\| \to 0$.

Proof or theorem 5.4 Since, from assumption SR1, the set Θ is compact, it follows that the set $\Theta - \{\theta_0\}$ can be covered by a finite number of balls $B(\lambda)$, $\lambda \neq \theta_0$, such that assumption SR2 holds for each one of them. We cannot apply Theorem 5.2 directly, because for each one of the balls we need to use a different normalizing sequence $\{A_T\}$. Nevertheless, the proof follows similar steps. Since there is a finite number of balls that cover $\Theta - \{\theta_0\}$, it is sufficient to focus on an open ball $B(\lambda)$, and to show that $\inf_{\theta \in B(\lambda)}\{S_T(\theta) - S_T(\theta_0)\} \to \infty$.

The least squares estimator $\hat{\theta}_T$ minimizes $S_T(\theta)$, and therefore it can equivalently be defined as the parameter value that minimizes $S_T(\theta) - S_T(\theta_0)$. We define (for every θ in $B(\lambda)$, and also θ_0)

$$L_T(\omega, \theta) := \frac{1}{A_T}\{S_T(\theta) - S_T(\theta_0)\},$$

where A_T is defined in equation (5.6). Then for $\theta \in B(\lambda)$

$$L_T^{**}(\omega, \theta) - L_T^{**}(\omega, \theta_0) = \frac{1}{A_T}\sum_{t=1}^{T}\{f_t(\theta) - f_t(\theta_0)\}^2 \geq 1,$$

which means that $\theta_T^{**} = \theta_0$. Since $A_T \to \infty$, it remains to show that $\mathbf{P}$-a.s.,

$$\sup_{\theta \in B(\lambda)} \frac{1}{A_T}\left|L_T(\omega, \theta) - L_T^{**}(\omega, \theta)\right| = o(1),$$

or more specifically that $\mathbf{P}$-a.s.

$$\sup_{\theta \in B(\lambda)} \frac{1}{A_T}\left|\sum_{t=1}^{T} \epsilon_t \{f_t(\theta) - f_t(\theta_0)\}\right| = o(1),$$

which calls for an application of Theorem 5.3. It is sufficient to verify only the assumptions U7 and U4, since assumption U1 follows immediately from SR1.

Assumption U7 holds since

$$\sum_{t=1}^{T} E_{t-1}\{\sup_{\theta\in B(\lambda)} \epsilon_t^2 \{f_t(\theta) - f_t(\theta_0)\}^2\} = \sum_{t=1}^{T} E_{t-1}(\epsilon_t^2) \sup_{\theta\in B(\lambda)} \{f_t(\theta) - f_t(\theta_0)\}^2 = O(A_T^{\rho\lambda}),$$

using (5.5) and (5.7). The last assumption we need to verify is U4. Now, for all $\theta_1, \theta_2 \in B(\lambda)$ we have

$$\left|\epsilon_t \{f_t(\theta_1) - f_t(\theta_0)\} - \epsilon_t \{f_t(\theta_2) - f_t(\theta_0)\}\right| = \left|\epsilon_t \{f_t(\theta_1) - f_t(\theta_2)\}\right| \leq |\epsilon_t|\, h(\|\theta_1 - \theta_2\|)\, M_t(\lambda),$$

where the last inequality follows from (5.8). Since $\sup_t E_{t-1}|\epsilon_t|^2 < \infty$, then $\sup_t E_{t-1}|\epsilon_t| < \infty$, which together with (5.9) imply that condition U4 holds. □

Chapter 6

Conclusions and Further Research

Prequential Statistical Forecasting System can be viewed from two different perspectives. First, they can be seen as purely predictive tools. For example Statistical Forecasting Systems are rules for probability forecasting, and Point Prediction Systems are point prediction rules. But, a Statistical Forecasting System can also be seen as an inferential tool. The prequential point of view considers every statistical model as a human attempt to explain nature, whose validity is to be assessed by the quality of the forecasts it produces. A Statistical Forecasting System can then been seen as a replacement of the statistical model which can be used to assess its validity, and to compare it with another model. In either case, either for optimal prediction or for inferential purposes, the property of efficiency of a forecasting system (which holds for a Bayesian forecasting system) is of great importance. In particular, prequential model selection based on efficient SFS's leads to consistent model selection. It was interesting therefore to study under what conditions efficient non-Bayesian SFS's exist.

In this thesis we showed that for regular models, under suitable conditions, plug-in SFS's can be efficient although they do not incorporate the parameter

uncertainty in their predictive distributions. We also demonstrated that plug-in SFS's can be inefficient, especially in cases where the information from the data grows fast.

There still remain some interesting open problems. An important question is how close is a prequentially efficient SFS to each $\mathbf{P}_\theta$ in $\mathcal{P}$. We might call a SFS $\mathbf{Q}$ *consistent* if $H(\mathbf{P}_{t,\theta}, \mathbf{Q}_t) \to 0$, under $\mathbf{P}_\theta$, for almost all $\theta \in \Theta$. According to this definition, efficiency is not a stronger property than consistency, since as it was discussed in Example 3.9, when $x'_{t+1}(X'_tX_t)^{-1}x_{t+1}$ does not converge to 0, then we can have an efficient SFS (the BFS) which is not consistent (since the predictive variance never converges to the true one). However, we conjecture (but have not as yet shown) that, whenever there does exist a consistent SFS, then any efficient SFS will be consistent (although the converse is of course false).

Another interesting direction for future research is the study of the notion of prequential efficiency applied to non-parametric families of sampling distributions. In that case Θ is not a Euclidean space, and care is now needed with the interpretation of "for almost all θ". For example we may be able to define efficient density estimators in non-parametric density estimation.

Also of great interest is the case where we have a sequence of parametric families $\mathcal{P}_t$, of increasing dimension, which approximate some "large" limiting model $\mathcal{P}_\infty$. When is a method, efficient for $\bigcup_t \mathcal{P}_t$, also efficient for $\mathcal{P}_\infty$? If not so, how well can one do with methods based on $\bigcup_t \mathcal{P}_t$?

In Chapter 4, we presented a new notion of efficiency for sequential point predictions, based on asymptotic empirical performance. We have shown that efficient predictors exist for general parametric families, under the weak and natural condition that their predictive variance stay bounded. The definition is applicable to linear and non-linear predictors, and to ergodic and non-ergodic models (Basawa and Scott, 1983). We showed that Bayesian Point Prediction Systems are efficient, when their predictive variances stay bounded, and presented sufficient conditions

for the efficiency of plug-in PPS's.

It would be interesting and important to develop sufficient conditions under which the predictive variances of Bayesian PPS's stay bounded almost surely, especially for non-ergodic models. One important special case is the autoregressive model with some of the roots inside and some outside the unit circle. Also it may be possible to weaken our conditions for efficiency of a plug-in PPS, especially for maximum likelihood estimators.

The notions of efficiency we have discussed are limited to probability forecasting and point prediction, using the squared prediction error as the loss function in the latter case. It would clearly be of interest to develop similar notions of efficiency for more general loss functions.

In Chapter 5, we showed how the notion of consistency under misspecification can be studied using a predictive point of view. Our results suggest that, under suitable conditions, the estimator based on the minimization of some statistical criterion that measures predictive performance converges to the parameter value that indexes the model that issues the best one step ahead predictions. This "best" model can vary for different sequences, and in order to overcome this difficulty we adopted a martingale framework, and proved a martingale version of the uniform law of large numbers. These results of course are also applicable in the case where our model is not misspecified, and may lead to some weakening of the standard conditions used for establishing consistency of extremum estimators. We have showed how this can be achieved for least squares estimators in nonlinear stochastic regression models.

References

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, **62**, 547–554.

Anderson, T. W. and Taylor, J. (1979). Strong consistency of least squares estimators in dynamic models. *Annals of Statistics*, **7**, 484–489.

Andrews, D. W. K. (1987). Consistency in nonlinear econometric models : A generic uniform law of large numbers. *Econometrica*, **55**, 1465–1471.

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

Basawa, I. and Scott, D. (1983). *Asymptotic Optimal Inference for non-ergodic Models.* Springer Verlag.

Basu, A. and Harris, I. (1994). Robust predictive distributions for exponential families. *Biometrika*, **81**, 790–794.

Blackwell, D. and Dubins, L. E. (1962). Merging of opinions with increasing information. *Annals of Mathematical Statistics*, **33**, 882–886.

Caines, P. E. (1988). *Linear Stochastic Systems.* Wiley Series in Probability and Mathematical Statistics.

Cencov, N. N. (1981). *Statistical Decision Rules and Optimal Inference.* Addison-Wesley.

Christopeit, N. and Helmes, K. (1980). Strong consistency of least squares estimators in linear regression models. *Annals of Statistics*, **8**, 778–788.

Clarke, B. S. and Barron, A. R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on information theory*, **36**, 453–471.

Clarke, B. S. and Barron, A. R. (1994). Jeffreys' prior is asympotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, **41**, 37–60.

Clarke, B. (1989). *Asymptotic Cumulative Risk and Bayes Risk under Entropy Loss with Applications.* Ph.D. thesis, University of Illinois.

Crowder, M. (1988). Asymptotic expansions of posterior expectations, distributions and densities for stochastic processes. *Ann. Inst. Statist. Math.*, **40**, 297–309.

Davis, M. H. A. and Vinter, R. B. (1985). *Stochastic Modelling and Control.* Chapman and Hall.

Dawid, A. P. (1984). Statistical theory. The prequential approach (with Discussion). *Journal of Royal Statistical Society, Series A*, **147**, 278–292.

Dawid, A. P. (1986). Probability forecasting. *Encyclopedia of Statistical Sciences*, **7**, 210–218.

Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference. *Journal of Royal Statistical Society, Series B*, **53**, 79–109.

Dawid, A. P. (1992a). Prequential analysis, stochastic complexity and Bayesian inference. *Bayesian Statistics 4*, 115–125.

Dawid, A. P. (1992b). Prequential data analysis. In Ghosh, M. and Pathak, P. (Eds.), *Current Issues in Statistical Inference : Essays in Honor of D. Basu.* IMS Lecture Notes-Monograph Series, Vol. 17, pp. 113–126.

Dawid, A. P. (1997). Prequential analysis. *Encyclopedia of Statistical Sciences Update Volume 1,* Samuel Kotz, Editor-in-Chief, 464–469.

El-Sayyad, G. M., Samiuddin, M., and Al-Harbey, A. A. (1989). On parametric density estimation. *Biometrica*, **76**, 343–348.

Gallant, A. R. and White, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models.* Basil Blackwell.

Gourieroux, C. and Monfort, A. (1995). *Statistics and Econometric Models. Volumes One and Two.* Cambridge University Press.

Harris, I. R. (1989). Predictive fit for natural exponential families. *Biometrika*, **76**, 675–684.

Ibragimov, I. A. and Hasminskii, R. Z. (1973). On the information in a sample about a parameter. *Second International Symposium on Information Theory* Akademiai, Kiado, Budapest, 295–309.

Ibragimov, I. A. and Hasminskii, R. Z. (1980). *Statistical Estimation : Asymptotic Theory.* Springer-Verlag.

Johnson, R. A. (1970). Asymptotic expansions associated with posterior distributions. *Annals of Mathematical Statistics*, **41**, 851–864.

Kabanov, Y. M., Liptser, R., and Shiryayev, A. N. (1978). Absolute continuity and singularity of locally absolute continuous probability distributions. *Math. USSR Sb.*, **35**, 631–680.

Kuboki, H. (1993). Inferential distributions for non-Bayesian predictive fit. *Ann. Inst. Statist. Math.*, **45**, 567–578.

Lai, T. and Wei, C. (1982). Least squares estimates in stochastic regression models with appications to identification and control systems. *Annals of Statistics*, **10**, 154–166.

Lai, T. and Wei, C. (1983). Asymptotic properties of general autoregressive models and strong consistency of least squares estimates of their parameters. *Journal of Multivariate Analysis*, **13**, 1–23.

Lai, T. L. (1994). Asymptotic properties of nonlinear least squares estimates in stochastic regression models. *Annals of Statistics*, **22**, 1917–1930.

Phillips, P. C. B. and Ploberger, W. (1994). Posterior odds testing for a unit root with data-based model selection. *Econometric Theory*, **10**, 774–808.

Phillips, P. C. B. (1996). Econometric model determination. *Econometrica*, **64**, 763–812.

Qian, G., Gabor, G., and Gupta. R. P. (1996). Generalised linear model selection by the predictive least quasi-deviance criterion. *Biometrika*, **83**, 41–54.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics*, **14**, 1080–1100.

Rissanen, J. (1987). Stochastic complexity. *Journal of Royal Statistical Society, Series B*, **49**, 223–239.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* Singapore : World Scientific.

Seillier-Moiseiwitsch, F., Sweeting, T. J., and Dawid, A. P. (1992). Prequential tests of model fit. *Scand. J. Statist.*, **19**, 45–60.

Shiryayev, A. (1996). *Probability (Second Edition).* Springer-Verlag.

Wei, C. (1987). Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. *Annals of Statistics*, **15**, 1667–1682.

White, H. and Wooldridge, J. (1991). Sieve estimation with dependent observations. In Barnett, W., Powell, J., and Tauchen, G. (Eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, pp. 459–493. New York : Cambridge University Press.

White, H. (1994). *Estimation, Inference and Specification Analysis.* Cambridge University Press.

Wu, C. F. (1981). Asymptotic theory of nonlinear least squares estimation. *Annals of Statistics*, **9**, 501–513.