Consistency in Misspecified Models

Kostas Skouras and A. Philip Dawid
Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK.

Abstract

We discuss the issue of consistency of extremum estimators for possibly misspecified models and present some new results which are applicable in cases where there is strong dependence between successive observations. Our approach extends that of White(1992) and is based on a martingale version of the law of large numbers, which we prove. We interpret our results using a predictivist point of view.

Keywords: Consistency; Extremum estimators; M-estimators; Misspecification; Prediction; Martingale Uniform Law of Large Numbers.

1 Introduction

Many problems in statistical modelling can be described as follows. A sequence of quantities \( Y = (Y_1, Y_2, \ldots) \), generated by an unknown data-generating process \( P \), is to be observed, and a statistical model \( M = \{ M(\theta) : \theta \in \Theta \} \) is proposed as a suitable description of some interesting properties of \( P \). The model \( M \) could, for example, be a class of probability distributions, a semi-parametric model describing only the means and variances of \( (Y_i) \), or a regression model describing a relationship between different components of \( Y \). Each constituent \( M(\theta) \in M \) will be referred to as a description.

In order to study the properties of different statistical methods, the assumption is usually made that there is a value \( \theta_0 \in \Theta \) such that \( M(\theta_0) \) is a ‘true’ description, in the sense that it describes the properties of interest accurately. In some situations this may
be a reasonable assumption to make, but in general we cannot expect our model to have captured all properties and relationships among the observed data, which may be very complex. The best we can hope for is that some member of \( \mathcal{M} \) be a good description, in some approximate sense which needs to be specified further, of the data generating process.

The purpose of statistical modelling and analysis based on the assumption of the existence of a true description within our model \( \mathcal{M} \) is well defined: we should try to identify, or make inferences about, that true description. But when \( \mathcal{M} \) is a misspecified model the purpose of modelling is less clear, as are the effects of misspecification on inferential procedures. In this case one needs to think carefully about the purpose, usefulness and limitations of using the model \( \mathcal{M} \).

As a simple example, suppose we observe a sequence of observations \( (Y_i) \), arising from a data-generating process under which they are independent and identically distributed with finite mean \( \theta \). We entertain a model \( \mathcal{M} \), consisting of descriptions \( M(\theta) \) under which the \( (Y_i) \) are independent and identically distributed Normal with mean \( \theta \) and variance 1. This family \( \mathcal{M} \) may not include the true data-generating process, since this may not be a Normal distribution, or may have variance different from 1. If we estimate \( \theta \) by the maximum likelihood estimator \( \hat{\theta}_T = (1/T) \sum_{t=1}^{T} Y_t \) under \( \mathcal{M} \) then the estimator \( \hat{\theta}_T \) is (strongly) consistent regardless of the true data-generating process and its mean \( \theta \). This consistency property does not mean that we can discover the ‘true’ distribution of the data, but only that a specific property of this distribution, the mean, can be consistently estimated using our misspecified model. However, attempts to use the model for other inferential purposes, for example forming prediction intervals, may be unsuccessful.

It is therefore important, when we model data using possibly misspecified models, to have a clear understanding of the features of the data we wish to describe, to use a statistical methodology that identifies (at least for large samples) the description 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 paper we study the behaviour of a wide class of estimators under misspecification. In order to interpret our results, we adopt a predictivist point of view, which considers
a statistical model as a method of making statements about the observable quantities \((Y_i)\). These statements can be phrased as sequential forecasts and therefore each model can be seen as a predictive system, \(e.g\). it may be a probability forecasting system, a point prediction system. In considering a model as a predictive rule we do not restrict the applicability of our study, since most statistical models can be seen in this light. For example, any probability distribution for a data-sequence can be regarded as a probability forecasting system (Dawid, 1984). Using this predictivist approach, we thus now assume that the modeller’s aim is to identify that description in \(M\) which issues predictions which are ‘closest’ to the optimal predictions under the true data-generating process.

In order to formalize the above, we introduce the following mathematical framework. Assume that the sequence of variables \(Y = (Y_1, Y_2, \ldots)\) is defined on a complete filtered probability space \((\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})\), such that \(Y_t\) is \(\mathcal{F}_t\)-measurable. Define \(Y^t := (Y_1, \ldots, Y_t)\). The increasing sequence of \(\sigma\)-algebras \((\mathcal{F}_t)\) describes the information available at time \(t\); in the simplest case \(\mathcal{F}_t = \sigma(Y^t)\). We denote the conditional expectation \(E(\cdot | \mathcal{F}_t)\) by \(E_t(\cdot)\).

We attempt to describe the data using a model \(M = \{M(\theta) : \theta \in \Theta\}\), indexed by a parameter \(\theta\) taking values in a compact separable metric space \((\Theta, d)\). We let \(\Theta\) be a metric space in order to allow our results to be applicable in infinite-dimensional parameter spaces. The \(\sigma\)-algebra used to define measurability on \(\Theta\) is the Borel \(\sigma\)-algebra generated by the open sets of \(\Theta\).

Assume that, on observing \(Y_t\), the description \(M(\theta)\) is penalized by means of a \(\mathcal{F}_t\)-measurable loss function \(l_t(\theta)\), the loss depending on the forecast produced by \(M(\theta)\), on the value of \(Y_t\), and possibly on other information contained in \(\mathcal{F}_t\). The normalized cumulative loss is denoted by \(L_T(\theta) := (A_T)^{-1}\sum_{t=1}^{T} l_t(\theta)\), where \((A_T)\) is a suitable normalizing sequence. We discuss in the following sections how this sequence should be chosen. Then \(L_T(\theta)\) is a \(\mathcal{F}_T\)-measurable real function measuring the empirical predictive performance of the description \(M(\theta)\) up to time \(T\).

For a specific set of observed data \(y^T\) of \(Y^T\), a standard method of choosing a description from \(M\) is to prefer that corresponding to the value \(\hat{\theta}_T\) of \(\theta\) minimizing \(L_T(\theta)\). Such estimators are called extremum estimators (Gourieroux and Monfort, 1995) or M-
estimators (Potscher and Prucha, 1991), and include many different estimators commonly used in parametric and non-parametric setups, e.g. prediction error estimators, maximum and pseudo-maximum likelihood estimators, least squares estimators, generalized method of moments. 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 description is a predictive system, we can replace the notion of a ‘true’ description with that of a ‘best’ description, where by ‘best’ we mean the description in $\mathcal{M}$ which issues the best predictions under the true data-generating process P. Then, we would hope that the extremum estimator $\hat{\theta}_T$ would converge, under suitable conditions, to the best value of $\theta$ as the number of observations tends to infinity. This property defines what we mean by consistency under a misspecified model. It is important to appreciate that (unlike the case when the model is true) the ‘best’ description $M(\theta)$, and hence the consistency property, can depend on both the loss function and the data.

Our aim in this paper is to study the consistency of extremum estimators for possibly misspecified models. In §2 we review H. White’s theory of inference for misspecified models, and in §3 we show that this theory is not applicable in some interesting models. In §4 we generalize White’s approach to include such cases, and in §5 derive the properties of this extension. The key to these results is a new martingale uniform law of large numbers, which we develop in §6. In §7 we highlight some further issues relating to misspecification and consistency, using examples. All proofs are collected together in the Appendix.

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 this approach to the problem of consistency.

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

In order to prove that the difference between the estimator \( \hat{\theta}_T \) and the ‘best’ value \( \theta^*_T \) tends to zero, as \( T \) tends to infinity, the following two assumptions are introduced (Gallant and White, 1988; White, 1994).

**Assumption W1. [Asymptotic Identifiability]**

For all \( \epsilon > 0 \),

\[
\liminf_{T \to \infty} \left\{ \min_{\theta : d(\hat{\theta}_T, \theta) \geq \epsilon} L^*_T(\theta) - L^*_T(\theta^*_T) \right\} > 0.
\]

**Assumption W2. [Uniform Law of Large Numbers]**

The sequence \( L_T(\theta) - L^*_T(\theta) \) obeys the strong uniform law of large numbers (ULLN):

\[
\sup_{\theta \in \Theta} |L_T(\theta) - L^*_T(\theta)| \to 0 \quad \text{P-a.s.}
\]

The assumption W1 is used to make sure that the functions \( L^*_T(\theta) \) do not become flat around \( \theta^*_T \), 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 and Wooldridge, 1991; White, 1994).

**Theorem 1** Under assumptions W1 and W2, \( d(\hat{\theta}_T, \theta^*_T) \to 0 \), \( \text{P-a.s.} \).

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 \( \theta \) that labels the description in \( \mathcal{M} \) which issues the best predictions in terms of the overall expected predictive loss.
3 Counterexamples

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

Example 1 (Stochastic level Dawid, 1991) Let $(X_t)$, $t \geq 0$, be a sequence of independent and identically distributed Normal variables with zero mean and unit variance. Let $Y_t = X_0 + X_t$, $t \geq 1$. Our class of descriptions 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 $\theta$. If $L_T(\theta) = (1/T) \sum_{t=1}^{T} (Y_t - \theta)^2$, 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. \hfill \Box

Example 2 (Mixture of Distributions) A sequence of independent and identically distributed random variables $(Y_t)$, $t \geq 1$ will be observed, and our parametric family of descriptions $\mathcal{P} = \{\mathbf{P}_\theta, \theta = 1, \ldots, k\}$ consists of a finite number of probability distributions for $Y = (Y_1, Y_2, \ldots)$, which are mutually singular. Let $\mathbf{p}_T(\theta)$ denote the density (with respect to the Lebesgue measure) of the joint distribution for $Y^T$, and $\mathbf{p}_i(\theta)$ the conditional density of $Y_i$ given $Y^{i-1}$ under $\mathbf{P}_\theta$. Let $L_T(\theta) := (1/T) \log \mathbf{p}_T(\theta) = (1/T) \sum_{i=1}^{T} \{-\log \mathbf{p}_i(\theta)\}$. The estimator $\hat{\theta}_T$ is the maximum likelihood estimator of $\theta$.

When the true description belongs to $\mathcal{P}$, then the estimator $\hat{\theta}_T$ converges almost surely to the true value of $\theta$ and is consistent. Assume now that the true description does not lie in $\mathcal{P}$, but it is a mixture of the above descriptions, i.e. $\mathbf{P} = \sum_\theta a_\theta \mathbf{P}_\theta$, with $\sum_\theta a_\theta = 1$. Let $\mathbf{P}^T$ and $\mathbf{P}^T_\theta$ denote the restrictions of $\mathbf{P}$ and $\mathbf{P}_\theta$ to the first $T$ observations. Then, under $\mathbf{P}$, the expectation of $L_T(\theta)$ is minimized at the value $\theta_T^*$ which minimizes the Kullback-Leibler distance $K(\mathbf{P}^T, \mathbf{P}^T_\theta)$ (we assume for simplicity that $\theta_T^*$ is unique). The sequence of minimizers $(\theta_T^*)$ is deterministic, and if White’s result were applicable we would expect the estimator $\hat{\theta}_T$ to converge to $\theta_T^*$ with probability one under $\mathbf{P}$. But this is not the case,
since it can easily be seen that $\mathbb{P}\{\hat{\theta}_T \to \theta\} = a_\theta$. This is another case where the estimator $\hat{\theta}_T$ converges to a data dependent limit. \hfill \Box

Example 3 (Linear Stochastic Regression) The observed variables $(Y_i)$ are generated from the following description:

$$Y_i = m_i + \epsilon_i$$

where $(\epsilon_i)$ is a martingale difference sequence, with respect to an increasing sequence of sub-algebras $(\mathcal{F}_i)$, and thus $m_i$ is the conditional mean of $Y_i$ given the past, i.e. $E(\epsilon_i | \mathcal{F}_{i-1}) = 0$ and $m_i = E(Y_i | \mathcal{F}_{i-1})$. Suppose that we try to model $Y_i$ using a linear model

$$E(Y_i | \mathcal{F}_{i-1}) = \theta' x_i,$$

such that $\theta \in \mathbb{R}^p$, and the regressors $x_i \in \mathbb{R}^p$ are $\mathcal{F}_{i-1}$ measurable. When $m_i$ is not equal to $\theta' x_i$, our model is misspecified. Let $L_T(\theta) := \sum_{t=1}^T (Y_i - \theta' x_i)^2$. The estimator which minimizes $L_T(\theta)$ is the least squares estimator $\hat{\theta}_T = (\sum_{t=1}^T x_i x_i')^{-1} \sum_{t=1}^T x_i Y_i$.

In this example the expected value of $L_T(\theta)$ may not exist without further assumptions on the overall expectations of the stochastic regressors $(x_i)$. Even then, the value of $\theta$ which minimizes the overall expectation of $L_T(\theta)$ 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 $\{x_i\}$, as the following lemma shows.

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

$$\sup_t E(|\epsilon_t|^a | F_{i-1}) < \infty \quad \text{(for some } a > 2),$$

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 $\sum_{t=1}^T x_i x_i'$. Then, with probability one under $\mathbb{P}$, $\|\hat{\theta}_T - \theta_T^{**}\| \to 0$. 

7
Proof. The estimator \( \hat{\theta}_T \) is equal to:
\[
\hat{\theta}_T = \left( \sum_{t=1}^{T} x_t x'_t \right)^{-1} \sum_{t=1}^{T} x_t y_t = \left( \sum_{t=1}^{T} x_t x'_t \right)^{-1} \sum_{t=1}^{T} x_t (m_t + \epsilon_t) \\
= \theta_T^{**} + \left( \sum_{t=1}^{T} x_t x'_t \right)^{-1} \sum_{t=1}^{T} x_t \epsilon_t.
\]
Using the results in (Lai and Wei, 1982), it can be shown that the difference \( \| \hat{\theta}_T - \theta_T^{**} \| \)
converges to zero. \( \Box \)

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.

4 An alternative view of consistency

As we discussed in §1, in each example of the previous section the loss function \( L_T(\theta) \)
can be 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 \( \theta \) 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, different sequences of observed data 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
descriptions, 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 relate the behaviour of the loss function \( L_T(\theta) \) 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 \(L_T(\theta) = (1/A_T) \sum_{t=1}^T l_t(\theta)\), for some \(\mathcal{F}_t\)-measurable functions \(l_t(\theta)\), then it may be more relevant to try to compare \(\hat{\theta}_T\) with the sequence \((\theta^{*}_T)\) of minimizers of the function

\[L^*_T(\theta) := (1/A_T) \sum_{t=1}^T E_{t-1}\{l_t(\theta)\},\]

instead of

\[L^+_T(\theta) = E\{1/A_T \sum_{t=1}^T l_t(\theta)\},\]

since the conditional expectation \(E_{t-1}\{l_t(\theta)\}\) is the conditional predictive risk, based on all observations up to time \(t-1\), whereas the overall expectation \(E\{l_t(\theta)\}\) is the unconditional predictive risk which does not take into account the observed data up to time \(t-1\). This approach also allows us to use a non-deterministic sequence \(A_T\) in the denominator, which 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 1-3, and give a natural interpretation to the limiting behaviour of \(\hat{\theta}_T\).

Example 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 conditional 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 then \(|\hat{\theta}_T - \theta^{*}_T| \to 0\). \(\square\)

Example 2 (contd.) Although under each \(P_\theta\), the observations \((Y_i)\) are independent identically distributed, under the mixture \(P\), the variables \((Y_i)\) are exchangeable, but not independent. Let \(\mathcal{F}_t = \sigma(Y_1, Y_2, \ldots, Y_t)\), \(l_t(\theta) = -\log p_t(\theta)\) and observe that \(L_T(\theta) = (1/T) \sum_{t=1}^T l_t(\theta)\). Denote by \(p_{t+1}\) the conditional density of \(Y_{t+1}\) given \(Y^t\) under \(P\). Then, it is well known that \(p_{t+1} = \sum_\theta a_\theta(\theta)p_{t+1}(\theta)\), where \(\{a_\theta(\theta)\}\) is the posterior distribution of \(\theta\) given the observations \(Y^t\). Since the distributions \(P_\theta\) are singular, then \(P\)-a.s. the posterior probability \(a_\theta(\hat{\theta}_t)\) converges to one. This implies that the Kullback
distance \( K(\mathbf{p}_{t+1}, \mathbf{p}_t(\hat{\theta}_t)) \) converges to 0, and for all \( \theta \neq \lim_t \hat{\theta}_t \), it stays positive, i.e. \( \liminf_t K(\mathbf{p}_{t+1}, \mathbf{p}_t(\theta)) > 0 \).

The function \( L_T^*(\theta) = (1/T) \sum_{t=1}^T E_t\{l_t(\theta)\} \) 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 3 (contd).** Observe that \( \theta^*_T = \left( \sum_{t=1}^T x_t x_t' \right)^{-1} \sum_{t=1}^T x_t y_t \) is the value of \( \theta \) that minimizes \( L_T^*(\theta) = \sum_{t=1}^T E_t\{(Y_t - \theta' x_t)^2\} \). □

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

## 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 establish the existence and measurability of \( \hat{\theta}_T \) and \( \theta^*_T \).

### 5.1 Existence

We introduce the following assumptions:

**Assumption E1.** The metric space \((\Theta, d)\) is compact and separable.

**Assumption E2.** (a) For every \( t \) and each \( \theta \) in \( \Theta \), the functions \( l_t(\theta) \) and \( E_t\{l_{t+1}(\theta)\} \) are \( \mathcal{F}_t \)-measurable, and continuous on \( \Theta \) almost surely, i.e. they are continuous for all \( \omega \) in an event \( F_t \in \mathcal{F}_t \) such that \( \mathbf{P}(F_t) = 1 \).

When assumptions E1 and E2 hold, then we can show that the functions \( \hat{\theta}_T \) and \( \theta^*_T \) exist almost surely and are measurable, using the following general lemma (Gallant and White, 1988; White, 1994; White and Wooldridge, 1991).
Lemma 2 Let $(\Omega, \mathcal{F})$ be a measurable space, and let $(\Theta, d)$ be a compact, separable metric space. Let $Q : \Omega \times \Theta \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$ be such that $Q(\cdot, \theta)$ is $\mathcal{F}$-measurable for each $\theta$ in $\Theta$, and $Q(\omega, \cdot)$ is continuous for all $\omega$ in an event $F \in \mathcal{F}$. Then there exists a function $\hat{\theta} : \Omega \rightarrow \Theta$ such that $\hat{\theta}$ is $\mathcal{F}$-measurable and for all $\omega$ in $F$

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

5.2 Consistency

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

Assumption C1. [Asymptotic Identifiability]

The function $L_T^{**}(\theta) := (1/AT) \sum_{t=1}^{T} E_{i-1} \{l_i(\theta)\}$ has a minimum, $P$-a.s., on $\Theta$ at $\theta^*_T$, for every $T$ sufficiently large. Let $\epsilon > 0$ and $B^c_T(\epsilon) := \{\theta \in \Theta : d(\theta, \theta^*_T) \geq \epsilon\}$. Then $P$-a.s.

$$\liminf_{T \to \infty} \left\{ \min_{\theta \in B^c_T(\epsilon)} L_T^{**}(\theta) - L_T^{**}(\theta^*_T) \right\} > 0. \quad (1)$$

Assumption C2. [Martingale Uniform Law of Large Numbers]

With probability one under $P$,

$$\sup_{\theta \in \Theta} |L_T(\theta) - L_T^{**}(\theta)| \to 0. \quad (2)$$

Theorem 2 Assume that conditions C1, C2 hold. Let $\hat{\theta}_T$ be an estimator that, $P$-a.s., minimizes $L_T(\theta)$ for all $T$ sufficiently large. Then, with probability one under $P$,

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

6 A Uniform Law of Large Numbers for Martingales

The main technical difference between White’s approach and ours 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.

We introduce now some extra assumptions and notation. Let $B(\theta, \rho)$ be the open ball around $\theta$ of radius $\rho$. Define:

$$
\overline{t}_t(\theta, \rho) = \sup_{s \in B(\theta, \rho)} l_t(s) \text{ and } \underline{t}_t(\theta, \rho) = \inf_{s \in B(\theta, \rho)} l_t(s).
$$

Let $A_T$ be a predictable increasing sequence with $A_1 \geq 1$ and $\lim_T A_T = \infty$. A sequence of ${\mathcal{F}_t}$-measurable 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 we need to use a predictable, and not a deterministic, sequence in the denominator.

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

**Assumption U1.** The metric space $(\Theta, d)$ 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 $(\overline{t}_t(\theta, \rho))$ and $(\underline{t}_t(\theta, \rho))$ satisfy pointwise strong martingale LLN’s (with common denominator an increasing predictable sequence $A_T$).

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

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

Although our assumptions are similar to Andrews’s (1987) assumptions, they are weaker because the use of conditional, instead of unconditional, expectations allow us to apply the result to cases where the sequence of functions $(l_t(\theta))$ may be strongly dependent.
Theorem 3 (Martingale ULLN) If Assumptions U1-U3 hold, then P-a.s.,

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

In many cases assumptions U1-U3 may be difficult to verify directly, and then we can use the following assumptions:

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

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

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, $B_t(\cdot)$ and $h(\cdot)$ may depend on $\theta$.

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

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

Assumption U6. With P probability one,

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

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

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

13
Lemma 3 The following hold:

(a) Assumption $U_4$ implies Assumption $U_3$.

(b) Assumption $U_5$ implies Assumption $U_4$.

(c) Assumption $U_6$ implies Assumption $U_2$.

(d) Assumption $U_7$ implies Assumption $U_2$.

7 Examples

In this section we present some examples in order to highlight some points on consistency for misspecified models.

Example 4 (AR(1) model) Let $Y_0 = 0$ and assume that

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

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

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

and estimate $\theta$ 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 $|\theta_0| < 1$ the true description is stationary and $\lim_T \hat{\theta}_T = \lim_T \theta^*_T = \lim_T \theta^{**}_T = 0$. When $|\theta_0| \geq 1$ the true description is not stationary and $\bar{Y}_T$ does not converge to $\theta^*_T$. But, $\hat{\theta}_T - \theta^{**}_T = (1/T) \sum_{t=1}^T \epsilon_t$ and, although $|\hat{\theta}_T - \theta^{**}_T| \not\rightarrow 0$, $|\hat{\theta}_T - \theta^{**}_T| \rightarrow 0$. 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 approximations to the true description, and therefore one should be careful to choose the appropriate loss function for his/her decision/prediction problem.
Example 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 $(\epsilon_t)$ 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. The first is to estimate $\theta$ using least squares, and then to substitute $\theta$ in the formula $E(Y_{T+2} | Y^T) = \theta^2 Y_T$ with the least squares estimator, and the second is to estimate $\phi = \theta^2$ directly by minimizing the 2-step ahead prediction errors $\sum_{t=3}^{T} (Y_t - \phi Y_{t-2})^2$.

Using Lemma 1, we can show that by minimizing the one step ahead prediction errors $\sum_{t=2}^{T} (Y_t - \theta Y_{t-1})^2$, 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 have that if we use the second method with an unrestricted value for $\phi$, 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 $\theta$, and consequently different predictions. It is easy to show 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 $\theta$ 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 $\theta$ 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 $\theta$ 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 $\theta$, in a misspecified model different loss functions give different
approximations to the true description. We should be careful therefore first to specify the decision problem we want to solve, and then to estimate \( \theta \).

The next example shows that, under misspecification, estimation consistency and prediction consistency are completely different aims.

**Example 6 (Error in variables)** Assume that the variables \( (Y_t) \) are generated from the model:

\[
Y_t = \theta_0 x_t + \epsilon_t,
\]

where \( (\epsilon_t) \) is a sequence of independent and identically distributed 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 \( \theta \) 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 1 we have \( |\hat{\theta}_T - \theta_T^{**}| \to 0 \). Since

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

we see that \( \hat{\theta}_T \) is estimation inconsistent, in the sense that it does not converges to the value \( \theta_0 \), but it is prediction consistent since it converges to the value of \( \theta \) that issues asymptotically the best one step ahead predictions.
A Proofs

Proof of Theorem 2. The following events have all probability one under $\mathbb{P}$:

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

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

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

Also, for all $\omega \in F$ and $T > T_2(\omega, \delta(\epsilon))$, $|L_T(\theta_T^{**}) - L_T^{**}(\theta_T^{**})| < \delta(\epsilon)/2$, so that

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

and for all $\omega \in F$ and $T > T_3(\omega, \delta(\epsilon))$, $|L_T^{**}(\hat{\theta}_T) - L_T(\hat{\theta}_T)| < \delta(\epsilon)/2$.

Then,

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

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

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

Proof of Theorem 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 $T \geq T_1(\omega, \theta)$,

\[ \frac{1}{A_T} \sum_{t=1}^{T} E_{t-1} \{ \ell_t(\theta, \rho) - \ell_t(\theta, \rho) \} < \epsilon. \]
The collection of balls \( \{ B(\theta, \rho(\theta)), \theta \in \Theta \} \), is an open cover of the compact set \( \Theta \), and therefore has a finite subcover \( \{ B(\theta_j, \rho(\theta_j)) : j = 1, 2, \ldots, J \} \).

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 \( T > \max_j T_1(\omega, \theta_j) \)

\[
\frac{1}{AT} \sum_{t=1}^{T} [l_t(s) - E_{\ell t-1} \{l_t(s)\}] \leq \frac{1}{AT} \sum_{t=1}^{T} \left( \bar{F}_t(\theta_1, \rho(\theta_1)) - E_{\ell t-1} \left[ \bar{F}_t(\theta_1, \rho(\theta_1)) \right] \right)
\]

\[
\leq \frac{1}{AT} \sum_{t=1}^{T} \left( \bar{F}_t(\theta_1, \rho(\theta_1)) - E_{\ell t-1} \left[ \bar{F}_t(\theta_1, \rho(\theta_1)) \right] \right) + \epsilon
\]

and

\[
\frac{1}{AT} \sum_{t=1}^{T} [l_t(s) - E_{\ell t-1} \{l_t(s)\}] \geq \frac{1}{AT} \sum_{t=1}^{T} \left( \bar{F}_t(\theta_1, \rho(\theta_1)) - E_{\ell t-1} \left[ \bar{F}_t(\theta_1, \rho(\theta_1)) \right] \right)
\]

\[
\geq \frac{1}{AT} \sum_{t=1}^{T} \left( \bar{F}_t(\theta_1, \rho(\theta_1)) - E_{\ell t-1} \left[ \bar{F}_t(\theta_1, \rho(\theta_1)) \right] \right) - \epsilon
\]

Then for every \( \omega \in F_0 \) and \( \theta \in \Theta \),

\[
\min_{j \leq J} \frac{1}{AT} \sum_{t=1}^{T} \left( \bar{F}_t(\theta_j, \rho(\theta_j)) - E_{\ell t-1} \left[ \bar{F}_t(\theta_j, \rho(\theta_j)) \right] \right) - \epsilon
\]

\[
\leq \frac{1}{AT} \sum_{t=1}^{T} [l_t(\theta) - E_{\ell t-1} \{l_t(\theta)\}]
\]

\[
\leq \max_{j \leq J} \frac{1}{AT} \sum_{t=1}^{T} \left( \bar{F}_t(\theta_j, \rho(\theta_j)) - E_{\ell t-1} \left[ \bar{F}_t(\theta_j, \rho(\theta_j)) \right] \right) + \epsilon.
\]

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

**Proof of Lemma 3.**

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

$$
\limsup_{\rho \to 0} \frac{1}{A_T} \sum_{t=1}^{T} E_{t-1} \{ \overline{l}_t(\theta, \rho) - l_t(\theta, \rho) \} \\
\leq \limsup_{\rho \to 0} \frac{1}{A_T} \sum_{t=1}^{T} E_{t-1} \{ \overline{\|l_t(\theta, \rho) - l_t(\theta, \rho)\|} + \|l_t(\theta, \rho)\| \} \\
\leq 2 \lim_{\rho \to 0} h(\rho) \sup_{T \geq 1} \frac{1}{A_T} \sum_{t=1}^{T} E_{t-1} \{ B_t(Y_t) \} = 0.
$$

$\Box$

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

In order to establish that U5 implies U4, we can use the mean value theorem to show that, $\mathbf{P}$-a.s.,

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

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

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

we get U4. $\Box$

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

For every $\theta$ and $\rho$ small enough we have,

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

and therefore

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

Thus,

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

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

$$
\frac{1}{A_T} \sum_{t=1}^{T} \overline{l}_t(\theta, \rho) - E_{t-1} \{ \overline{l}_t(\theta, \rho) \} \rightarrow 0.
$$
The same argument can be used for $l_d(\theta, \rho)$ and $l_1(\theta)$.  

(d) Assumption U7 implies Assumption U2.

We will use a similar method as in (c). For every $\theta$ and $\rho$ small enough we have,

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

and therefore

$$\sum_{t=1}^{T} E_{t-1}\{\sup_{\theta \in \Theta} |l_t(\theta)|^2\} \geq \sum_{t=1}^{T} V_{t-1}\{\bar{l}_t(\theta, \rho)\}, \quad (3)$$

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

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

Using (3) and the fact that

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

we get

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

The same argument can be used for $l_d(\theta, \rho)$ and $l_1(\theta).$  

References


