Identification Robust Predictive Ability Testing

(Job Market Paper 2)
Most recent version: here

Andrea A. Naghi *

University of Warwick

October 2016

Abstract

The paper analyzes the predictive accuracy evaluation of models that are strongly identified in some part of the parameter space but non-identified or weakly identified in another part of the parameter space. We show that when comparing the predictive ability of models that might be affected by identification deficiencies, when the parameter estimation error is negligible, the null distribution of out-of-sample predictive ability tests is not well approximated by the standard normal distribution. As a result, employing a standard (strong) identification critical value can lead to misleading inference. We propose methods to make the out-of-sample predictive ability tests robust to identification loss. These methods use a different critical value than the standard one and include: a least-favorable critical value and a data dependent critical value. In settings where the parameter error is non-negligible, the paper shows that the asymptotic distribution of the predictive ability test in West(1996) is standard, even when one allows for the model(s) to be only semi-strongly identified.

*Department of Economics, University of Warwick, email: a.a.naghi@warwick.ac.uk. I would like to thank Valentina Corradi for her guidance, helpful comments and advice. I also benefited from insightful discussions with Ana Galvao and Jeremy Smith. Feedback and helpful suggestions from Clement De Chaisemartin, Bo Honore, Arthur Lewbel and Giovanni Ricco helped improve the paper. Financial support from the ESRC and RES is gratefully acknowledged.
Keywords: predictive ability, weak identification, semi-strong identification, identification loss, robust inference

J.E.L Classification: C12, C15, C22, C53
1 Introduction

The paper considers the predictive ability evaluation of models affected by identification deficiencies. Comparing the predictive ability of models where identification might fail, has not yet been examined in the literature, even though many models from empirical predictive evaluation studies can suffer from identification problems. Consider for example the Smooth Transition Autoregressive (STAR) models or the simple ARMA(1,1) model with close to equal autoregressive (AR) and moving-average (MA) parameters. A range of other examples of models that satisfy our identification setup is given in Section 2.2.

The class of models we are concerned with, are models where the parameter of interest $\theta$, is of the form $\theta = (\beta, \zeta, \pi)$, where $\pi$ is identified if and only if $\beta \neq 0$. The parameter $\zeta$ is not related to the identification of $\pi$, while $\psi = (\beta, \zeta)$ is always identified. In our setup, the potential source of identification deficiency is an intrinsic characteristic of the model. For example, in the STAR models, one can lose identification in part of the parameter space, because the parameter that measures the slope of the transition and the parameter that gives the location of the transition\(^1\) are not identified when the coefficient on the nonlinear regressor is zero.

Given the structure of the class of models under consideration, when $\beta = 0$, a criterion function\(^2\), used to estimate the parameters of the model, does not depend on $\pi$. When $\beta$ is close to 0, the criterion function is flat with respect to $\pi$. As a result, the second derivative matrix of the criterion function is singular or near-singular causing the failure of standard asymptotic approximations, since they involve the inverse of this matrix.

The magnitude of $\|\beta\|$ determines the identification strength of the model (or equivalently the parameter $\pi$)\(^3\), which as discussed below can be: non-identification, weak identification, semi-strong identification\(^4\) or strong identification. In an in-sample estimation framework, Andrews and Cheng (2012), Andrews and Cheng (2013) and Andrews and Cheng (2014), hereafter AC2012, AC2013 and AC2014 respectively, show that in this class of models, estimators can be inconsistent or consistent with different rates of

\(^1\)Note that $\beta$, $\pi$ and $\zeta$ can be scalars or vectors.
\(^2\)This criterion function can be: maximum likelihood criterion function, least-square criterion function, GMM criterion function, minimum distance criterion function and others.
\(^3\)Throughout the paper, when $\beta$ is a vector, $\|\cdot\|$ denotes the Euclidean norm, when $\beta$ is scalar, $\|\cdot\|$ is replaced by the absolute value function.
\(^4\)Semi-strong identification is a technical construct, used to bridge the gap between weak and strong identification. See Section 2.1 for details.
convergence and have different asymptotic distributions, depending on the identification strength. The asymptotic null distribution of test statistics such as $t$, $QLR$ or $Wald$ \(^5\) is shown to be nonstandard when the model is non-identified or weakly identified. Once the parameter $\pi$ is at least semi-strongly identified, the asymptotic null distribution of these tests becomes standard.

In this paper, we analyze the effect of identification loss in an out-of-sample predictive ability evaluation framework. We show that in the considered class of models, under weak and non-identification, the finite-sample distributions of some of the estimators are far from the normal distribution. In our example using a STAR model, these finite sample distributions can be strongly bimodal or uniform. Due to these nonstandard distributions, the forecast errors generated from a model that is not identified are larger than the errors that would have been obtained if the model was (strongly) identified. This affects the null distribution of out-of-sample predictive ability tests, which are not well approximated by the standard normal distribution when one (or both) model(s) considered for predictive ability comparison are not (strongly) identified, in situations in which the parameter estimation error is negligible.

In a numerical example, we show that tests and confidence intervals that employ the standard normal critical value have size distortions and can lead to incorrect inference. The weaker is the identification strength, the lower are the coverage probabilities of confidence intervals (CI) based on the standard normal critical value.

We propose methods to make the out-of-sample predictive ability tests and CIs robust to identification deficiencies. These methods use a different critical value than the standard one and include: a least-favorable critical value and a data dependent critical value. The least favorable critical value is based on the least favorable possible distribution of the test statistic. For the data-dependent critical value, we use an identification category selection procedure to determine the identification strength of $\pi$ and adjust the critical value accordingly.

In settings where the parameter error is non-negligible, we show that the asymptotic distribution of the predictive ability test in West (1996) is standard even when we allow for $\pi$ to be only semi-strongly identified.\(^6\) Semi-strong identification does not affect the

\(^5\)The restrictions tested by these tests involve in general both $\psi$ and $\pi$.

\(^6\)Semi-strong identification is not covered by the results in West (1996), only the usual strong identification.
asymptotic distribution of the test statistic, even though it is known that under semi-
strong identification the estimator of \( \pi \) has a slower rate of convergence than the standard
(square-root of the sample size) rate.

**Related Literature.** The paper contributes to the strand of literature that focuses on
predictive ability evaluation of forecasting models. Predictive accuracy evaluation topics
already considered in the literature include among others: accounting for nested models
(Clark and McCracken (2001, 2005), McCracken (2007)), finite-sample predictive ability
(Giacomini and White (2006), Clark and McCracken (2009)), multiple model comparison
(White (2000), Hansen (2005), Corradi and Distaso (2011)), evaluation of conditional
quantile forecasts (Giacomini and Komunjer (2005)), tests for density forecasts (Corradi
and Swanson (2006), Corradi and Swanson (2006b), Amisano and Giacomini (2007)),
predictive ability evaluation in unstable environments (Giacomini and Rossi (2010), Rossi
and Sekhposyan (2013)), evaluation of factor-augmented models (Goncalves et al. (2015),
Fosten (2016)), allowing for latent target variables (Li and Patton (2013)). For surveys
on recent developments in predictive accuracy methodology see for example Corradi and

The focus of this paper is on predictive evaluation of models where the assumption
of (strong) identification fails. Recent papers such as AC2012, AC2013, AC2014 and
Cheng (2015) have analyzed the type of models considered in this paper, in an in-sample
estimation framework. AC2012 provide results for general extremum estimators. Under
high-level assumptions regarding the behavior of the estimator criterion function, they
establish consistency/lack-of-consistency of estimators and determine the asymptotic dis-
tributions of estimators under a range of drifting sequences of true distributions. The
properties of tests and confidence sets for parameters are analyzed under different identi-
fication categories. AC2013 and AC2014 provide a set of primitive conditions for the high
level assumptions in AC2012. AC2013 focus on models estimated by log-likelihood crite-
rian functions, while the focus of AC2014 is on moment condition models estimated by
GMM. The focus of Cheng (2015) is on mixed identification. While AC2012, AC2013 and
AC2014 treat a broader class of models with a single source of non-identification, Cheng
(2015) provides results for nonlinear regression models with multiple nonlinear regressors
and thus multiple sources of non/weak identification.

Related to the problem analyzed in the above-mentioned papers is the setting consid-
ered in Shi and Phillips (2012), where they allow for the possibility of weak identification.
in models characterized by a nonlinear cointegrating relationship. Weak identification in their model arises from the presence of a loading coefficient associated to the nonlinear function that might be close to zero. Andrews, Cheng and Guggenberger (2011) provide generic results used to convert asymptotic results under drifting sequences of parameters into results that hold uniformly over the parameter space. These results are useful in establishing the asymptotic size in a uniform sense of confidence sets and tests, and they can be applied, among others, to the class of models considered in this papers.

The paper is also broadly related to many other papers on weak and non-identification as well as papers on robust inference in weakly identified models. Consider for example the literature on robust inference with weakly identified nuisance parameters (see for e.g. Chaudhuri and Zivot (2011), Dufour and Taamouti (2005, 2007), Guggenberger et al. (2012)), or papers on weak identification in macroeconomic models (see for e.g. Andrews and Mikusheva (2015, 2016), Guerron-Quintana et al. (2013), Qu (2014)).

The paper is organized as follows. Section 2 introduces the class of models that we focus on, describes the identification problem and outlines the predictive ability evaluation framework. Section 3 provides numerical evidence on the implications of identification loss. Section 4 establishes asymptotic results under semi-strong identification. Section 5 proposes methods to construct robust tests and confidence intervals and applies them in a numerical example. Section 6 concludes. The Appendix provides proofs and additional numerical results.

2 Setup

This section presents the class of models that we consider in our predictive evaluation framework and its identification characteristics. We further give a couple of examples of models used in empirical studies that satisfy our identification setup and outline the predictive ability evaluation framework.
2.1 Class of Models and Identification Categories

Consider the class of regression models with additive nonlinearity:

\[ y_t = Z_t'\zeta + f(X_t, \pi)'\beta + u_t \]  

(1)

where \( y_t \in \mathbb{R} \), \( X_t \in \mathbb{R}^{d_X} \), \( Z_t \in \mathbb{R}^{d_Z} \) are observed variables, and \( u_t \in \mathbb{R} \) is an unobserved error term. The models belonging to (1) are nonlinear parametric regressions\(^7\), where \( f(X_t, \pi) \) is a smooth nonlinear function with unknown transformation parameter \( \pi \in \mathbb{R}^{d_\pi} \). The function \( f(X_t, \pi) \) is known up to the finite dimensional parameter \( \pi \). The loading coefficient \( \beta \in \mathbb{R}^{d_\beta} \) measures the importance of the nonlinear component, while \( \zeta \in \mathbb{R}^{d_\zeta} \) are the coefficients of some linear regressors in the model.

The inherent characteristic of the class of models in (1) is that \( \pi \) is identified if and only if \( \beta \neq 0 \). In fact, \( \|\beta\| \) determines the identification strength of \( \pi \). The parameter \( \beta \) can converge to 0 at various rates or, on the contrary, can be bounded away from 0. The parameter \( \zeta \) associated to the linear part of the regression is not related to the identification of \( \pi \), while \( \psi = (\beta, \zeta) \in \mathbb{R}^{d_\psi} \) is always identified.\(^8\) The parameters of interest in the model can be collected in a vector \( \theta = (\beta, \zeta, \pi) \in \mathbb{R}^{d_\theta} \).

Models of the type given in (1) are strongly identified at some points of the parameter space, but might be unidentified or weakly identified at other points of the parameter space. This feature requires non-standard inference techniques for this class of models. Moreover, the large sample properties of estimators, tests and confidence sets differ according to the identification properties of the model (see for e.g. AC2012, Cheng (2015)).

When \( \beta = 0 \), the criterion function used to estimate the parameters of the model, does not depend on \( \pi \), while when \( \beta \) is close to zero, the criterion function is flat with respect to \( \pi \). This means that the second derivative matrix of the criterion function is singular or near singular, causing the failure of standard asymptotic approximations, as they involve the inverse of this matrix.

In order to model the different identification strengths of \( \pi \), alternative asymptotic approximations along drifting sequences of true parameters need to be considered. To

\(^7\)The ARMA(1,1) model with equal autoregressive and moving-average coefficients presents similar identification issues as (1), but it is linear.

\(^8\)This setup also allows for cases where a model is re-parametrized to convert it into the framework considered here.
this end, suppose that the true value of the parameter is: \( \theta_T = (\beta_T, \zeta_T, \pi_T) \), where \( T \geq 1 \) indexes the sample size. As shown in AC2012 the behaviour of the extremum estimators and test statistics associated to models in (1), depends on the magnitude of \( \|\beta_T\| \) and varies across three categories of sequences \( \{\beta_T : T \geq 1\} \), which, following the terminology in AC2012, are:

**Category I (a):** when \( \beta_T = 0 \) \( \forall T \geq 1 \)

**Category I (b):** when \( \beta_T \neq 0 \) and \( T^{1/2}\beta_T \to b \in \mathbb{R}^{d_\beta} \)

**Category II:** when \( \beta_T \to 0 \) and \( T^{1/2}\|\beta_T\| \to \infty \)

**Category III:** when \( \beta_T \to \beta_0 \neq 0 \), \( \beta_0 \in \mathbb{R}^{d_\beta} \)

When the sequences \( \{\beta_T\} \) satisfy the conditions from category I (a), category I (b), category II and category III, the parameter \( \pi \) is, respectively, unidentified, weakly identified, semi-strongly identified and strongly identified. Note that when \( \pi \) is weakly identified, \( \beta_T \) converges to zero at the same rate as \( 1/\sqrt{T} \), while when \( \pi \) is semi-strongly identified, \( \beta_T \) converges to zero slower than \( 1/\sqrt{T} \).

As pointed out in AC2012, \( \theta = (\beta, \zeta, \pi) \) does not need to completely determine the distribution of the data. Thus, an additional parameter \( \phi \) can be introduced, which is an infinite dimensional nuisance parameter, such that \( \gamma = (\theta, \phi) \in \Gamma \) fully determines the distribution of the data.\(^9\) Therefore, in addition to the drifting sequences \( \{\beta_T : T \geq 1\} \), we can allow other parameters to change with the sample size and obtain uniform results not only over \( \beta \) but also over \( \gamma \).

Based on the identification categories described above, define the following sequences of true parameters \( \{\gamma_T\} \), where \( \{\gamma_T\} = \{(\theta_T, \phi_T)\} \):

1. \( \Gamma(\gamma_0) = \{\gamma_T : T \geq 1 : \gamma_T \to \gamma_0 \in \Gamma\} \)
2. \( \Gamma(\gamma_0, 0, b) = \{\gamma_T \in \Gamma(\gamma_0) : \beta_0 = 0 \text{ and } T^{1/2}\beta_T \to b \in (\mathbb{R} \cup \pm \infty)\} \)
3. \( \Gamma(\gamma_0, \infty, \omega_0) = \{\gamma_T \in \Gamma(\gamma_0) : T^{1/2}\beta_T \to \infty \text{ and } \beta_T/\|\beta_T\| \to \omega_0 \in \mathbb{R}^{d_\beta}\} \)

\(^9\)For example, in nonlinear regression models estimated by least squares, \( \theta \) indexes a finite-dimensional feature of the distribution of the errors (such as its variance), while \( \phi \) indexes the remaining characteristics of the distribution of the errors, which may be infinite-dimensional.
where \( \gamma_0 = (\beta_0, \zeta_0, \pi_0, \phi_0) \in \Gamma \) is the limit of the sequences \( \{\gamma_T\} \). In the notation above, the 0 in \( \Gamma(\gamma_0, 0, b) \) means \( \beta_0 = 0 \) while the \( \infty \) in \( \Gamma(\gamma_0, \infty, \omega_0) \) stands for \( T^{1/2}\|\beta_0\| \rightarrow \infty \).

The sequences \( \{\gamma_T\} \) in the set \( \Gamma(\gamma_0, 0, b) \), are sequences for which \( \{\beta_T\} \) is close to 0 and are in identification category I and II. The sequences \( \{\gamma_T\} \) in the set \( \Gamma(\gamma_0, \infty, \omega_0) \) are sequences for which \( \{\beta_T\} \) is more distant from 0, and are in identification category II and III. Note that both sets \( \Gamma(\gamma_0, 0, b) \) and \( \Gamma(\gamma_0, \infty, \omega_0) \) contain sequences \( \{\gamma_T\} \) from identification category II. What distinguishes these two sets is whether \( T^{1/2}\|\beta_T\| \rightarrow \|b\| \) with \( \|b\| < \infty \) or \( \|b\| = \infty \).10

### 2.2 Examples

The class of models considered in the paper includes many well known models which appear frequently in empirical studies.

**Example 1. Smooth Transition Autoregressive (STAR) Models**

The model can be written as follows:

\[
y_t = X_t'\zeta + X_t'\beta \cdot f(V_t, \pi) + u_t,
\]

with \( X_t = (1, y_{t-1}, \ldots, y_{t-p})' \) and \( V_t = y_{t-d} \) (2)

The function \( f(V_t, \pi) \) is known and the values of the parameters \( p \) and \( d \) are also assumed to be known, with \( 1 \leq d \leq p \). The literature usually considers two different forms for the transition function \( f(V_t, \pi) \in [0, 1] \): the logistic function

\[
f(V_t, \pi) = (1 + \exp [-\pi_1^*(V_t - \pi_2^*)])^{-1}
\]

or the exponential function

\[
f(V_t, \pi) = 1 - \exp [-\pi_1^*(V_t - \pi_2^*)^2]
\]

where \( \pi = (\pi_1^*, \pi_2^*)' \in \mathbb{R}^2 \). The parameter \( \pi_1^* > 0 \) measures the slope of the transition, while the parameter \( \pi_2^* \) measures the location of the transition. When \( \beta=0 \), the parameter \( \pi \) is not identified. The model can be estimated using an estimator criterion function such as the least square or maximum likelihood criterion function. The first derivatives of these

---

10Note that \( b \) indexes the magnitude of \( \beta \).
criterion functions with respect to $\pi$ are proportional to $\beta$, when $\beta$ is close to zero. Hence, the criterion functions are flat in the direction of $\pi$. For applications of this model see for e.g. Luukkonen et al. (1988), Terasvirta and Anderson (1992), Teräsvirta (1994), Ferrara, Marcellino and Mogliani (2015), among others.

**Example 2. Nonlinear Regression with Endogeneity**

Consider the following nonlinear regression model with endogenous regressors:

$$y_t = \beta \cdot f(X_{1t}, \pi) + X_{2t}'\zeta + u_t$$ with instruments $Z_t$ independent of $u_t$ (5)

where $X_{1t}$ is a nonlinear regressor, $X_{2t}$ is a linear regressor, and $Z_t$ are the instrumental variables, which are assumed to be strong. Identification failure in this model can arise due to the nonlinearity introduced by the function $f$. This model can be estimated using a GMM criterion function. In this example, the endogeneity of the regressors is not crucial for our identification problem. The reason to consider it here is to be able to give example of different extremum estimators that can be used. An application of this model is considered in Areosa et al. (2011), where they model inflation rate in Brazil and find strong support for a nonlinear specification of the Philips Curve. In their empirical study, the nonlinear function $f$ is a logistic function.

**Example 3. ARMA(1,1) model**

It is a well known result that common autoregressive (AR) and moving average (MA) roots lead to identification failure in the ARMA(1,1) model. In other words, the AR and MA parameters are not identified when they are equal. This occurs when the series is white noise. Consider the following ARMA(1,1) model:

$$y_t = (\pi + \beta)y_{t-1} + \epsilon_t - \pi\epsilon_{t-1}$$ (6)

when $\beta = 0$ the model becomes $y_t = \pi y_{t-1} + \epsilon_t - \pi\epsilon_{t-1}$ which is equivalent to $y_t = \epsilon_t$. Thus, when $\beta = 0$ the AR parameter $\pi + \beta$ and the MA parameter $\pi$ are not identified. Ansley and Newbold (1980) and Nelson and Startz (2007) show in simulations that AR and MA parameters that are close in value, lead to bias, variance and size problems. The ARMA(1,1) example fits our identification setup, even though there is no nonlinear function involved.

Other examples of models that satisfy our identification setup include: mixed data
sampling (MIDAS) regressions in empirical finance (see e.g. Ghysels, Sinko and Valkanov (2007)), autoregressive distributed lag models, continuous transition structural change models, continuous transition threshold autoregressive models (see e.g. Chan and Tsay (1998)), models with correlated random coefficients (see e.g. Andrews (2001)), nonlinear binary choice models, and probit models with endogeneity and possibly weak instruments (see e.g. Rivers and Vuong (1988)). The focus of this paper is on the subset of models that satisfy the identification setup described above and are also used in prediction.

Our framework does not cover all models that have identification deficiencies at some points in the parameter space. For example, models such as regime switching models, abrupt transition structural change models, and abrupt transition threshold autoregressive models, such as in Hansen (2000), Liu and Shao (2003), Elliott and Müller (2007, 2014), Qu and Perron (2007) are not considered here and are left for future research.

### 2.3 Predictive Evaluation Framework

We are interested in comparing the predictive ability of two models where one or both belong to the class described in Section 2.1, in situations where one or both models are affected by identification deficiencies. Here, we outline the predictive evaluation framework with notations corresponding to the case when both models belong to (1).

Consider thus the forecasting model 1:

\[ y_{t+h} = Z'_{1t} \zeta_1 + f_1(X_{1t}, \pi_1)' \beta_1 + u_{t+h} \]  

(7)

and a benchmark model, model 2:

\[ y_{t+h} = Z'_{2t} \zeta_2 + f_2(X_{2t}, \pi_2)' \beta_2 + \epsilon_{t+h} \]  

(8)

where \( y_t \in \mathbb{R}, X_{1t} \in \mathbb{R}^{d_{x_1}}, Z_{1t} \in \mathbb{R}^{d_{z_1}}, X_{2t} \in \mathbb{R}^{d_{x_2}}, Z_{2t} \in \mathbb{R}^{d_{z_2}} \) are observed variables, \( u_t \in \mathbb{R} \) and \( \epsilon_t \in \mathbb{R} \) are unobserved error terms and \( h \) is the forecast horizon.

Alternatively, the benchmark forecast could be also considered a non-model based forecast such as those obtained from different surveys. In this paper, we maintain a non-nested framework, i.e. we consider the case where \( W_{2t} = (X_{2t}, Z_{2t}) \) is not nested within

\[11\text{For other examples and references for models from this class, see AC2012.} \]
$W_t = (X_{1t}, Z_{1t})$.

In order to test the predictive accuracy of the model in (7) relative to the accuracy of the model in (8), we can formulate the null hypothesis of equal unconditional predictive ability as:

$$H_0 : E[g(u_{t+h}) - g(\epsilon_{t+h})] = 0$$

(9)

where $u_{t+h}$ and $\epsilon_{t+h}$ are the forecast errors of model (7) and model (8) and $g(\cdot)$ is some differentiable (or non-differentiable) loss function. This null hypothesis tests the equality of expected forecast error losses resulted from the two competing models. The alternative can be two sided (the two models do not perform equally well) or one-sided (one of the models performs better).

The forecast errors of model (7) and model (8) can be obtained by using a rolling, recursive or a fixed window estimation scheme. The notations in this paper correspond to the rolling estimation scheme, but recursive and fixed estimation cases could be also considered. Thus, the sample of size $T$ is divided into an in-sample part $R$ and an out-of-sample part $P$ and the models’ parameters are re-estimated for each rolling window of length $R$, indexed by $t$ with $R \leq t \leq T$.

Suppose the parameters of interest of the model in (7), $\theta_1 = (\zeta_1, \beta_1, \pi_1)$ are estimated by minimizing a criterion function $Q_R(\theta_1)$, that depends on the observables and the sample size $R$, over a parameter space $\Theta_1 \subseteq \mathbb{R}^{d_{\theta_1}}$. The criterion function is minimized successively over rolling windows of data. Analogously, $\theta_2$ is estimated by minimizing a criterion function $Q_R(\theta_2)$ over an optimization parameter space $\Theta_2 \subseteq \mathbb{R}^{d_{\theta_2}}$.

Given the parameter estimates, the forecast errors based on a rolling window of observations from model (7) and model (8) are given by:

$$\hat{u}_{t+h} = y_{t+h} - Z'_{1t}\hat{\zeta}_{1,R,t} - f(X_{1t}, \hat{\pi}_{1,R,t})'\hat{\beta}_{1,R,t} \quad \text{where } R \leq t \leq T$$

(10)

and

$$\hat{\epsilon}_{t+h} = y_{t+h} - Z'_{2t}\hat{\zeta}_{2,R,t} - f(X_{2t}, \hat{\pi}_{2,R,t})'\hat{\beta}_{2,R,t} \quad \text{where } R \leq t \leq T$$

(11)

Thus, an out-of-sample predictive ability test statistic such as the Diebold-Mariano-West
(DMW) statistic can be formulated as:

\[ S_P = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} (g(\hat{u}_{t+h}) - g(\hat{\epsilon}_{t+h})) \]  \hspace{1cm} (12)

and used to test the null given in (9).

The main difference between our setup and the classical DMW framework is the dependence of the test statistic on the estimates of the parameters \( \pi_1 \) and \( \pi_2 \), where \( \pi_1 \) and/or \( \pi_2 \) might not be strongly identified.

3 Numerical Evidence: Implications of Identification Loss

In this section we illustrate numerically the impact of identification issues, that arise naturally in the class of models described in Section 2, on the finite-sample densities of estimators, as well as on the finite-sample densities of out-of-sample tests and critical values.

3.1 Effect on Estimators

As an example of a model from the class (1), consider the following LSTAR(1):

\[ y_t = \zeta_1 + \zeta_2 y_{t-1} + \beta \cdot f(y_{t-1}, \pi) + u_t \]  \hspace{1cm} (13)

where \( f(x, \pi) = x(1 + \exp [-10(x - \pi)])^{-1} \) and \( u_t \sim N(0, 1) \).

In this simulation exercise, for illustration purposes we use the constant 10 for the parameter that measures the slope of the transition. The identification issues are already evident when only one parameter (the location of the transition) appears in the logistic function. Estimating a second parameter in the logistic function will only aggravate the identification problems.

Figures 1 to 4 provide the finite sample densities of the maximum likelihood estimators for \( \zeta_1, \zeta_2, \beta, \) and \( \pi, \) respectively, when the true value of \( \pi, \pi_0 = -1.5. \) The number of simulation repetitions is 500, and the sample size is \( T = 500. \) Each subplot gives the
densities of estimators for $b = [0, 2, 4, 10]$ where $b = \sqrt{T} \beta$ indexes the magnitude of $\beta$. The values of $b = [0, 2, 4, 10]$ correspond to $\beta = [0, 0.09, 0.18, 0.45]$. The true values of $\zeta_1$ and $\zeta_2$ are set to -1 and 0.5 respectively.

Figures 1 and 2 show that the finite-sample densities of the estimators of $\zeta_1$ and $\zeta_2$ are not far from a normal distribution for all values of $b$. In contrast, the results for the estimators of $\beta$ and $\pi$ are different, the finite-sample distributions being very far from a normal distribution for smaller values of $b$.

![Figure 1](image1.png)

**Figure 1:** Finite sample densities of the estimator of $\zeta_1$ when $\pi_0 = -1.5$.

![Figure 2](image2.png)

**Figure 2:** Finite sample densities of the estimator of $\zeta_2$ when $\pi_0 = -1.5$.

Figure 3 shows that the maximum likelihood estimator of $\beta$ has a clear bimodal distribution when $b = 0$ (non-identification of $\pi$), $b = 2$ and $b = 4$ (weak identification of $\pi$). It starts to resemble a normal distribution centered around the true value $\beta_0 = 0.45$ once $b$ is increased to 10. Regarding the distribution of the estimator of $\pi$, Figure 4
indicates that it is almost uniformly distributed on the interval [-4, -0.5], for a true value of \( \pi_0 = -1.5 \), when \( b = 0 \). When \( b = 2 \) it still resembles a uniform distribution, with a very slight build-up of mass that starts forming around the true value. The build-up of mass becomes more evident when \( b = 4 \), and the distribution of the maximum likelihood estimator of \( \pi \) is not far from a normal distribution centered around the true value once \( b = 10 \).

![Figure 3](image1.png)

**Figure 3:** Finite sample densities of the estimator of \( \beta \) when \( \pi_0 = -1.5 \).

![Figure 4](image2.png)

**Figure 4:** Finite sample densities of the estimator of \( \pi \) when \( \pi_0 = -1.5 \).

Similar results on the densities of estimators when \( \pi_0 = -3 \) are reported in Figures 9 to 12, in the Appendix.
3.2 Effect on Out-of-Sample Tests

This section focuses on the case in which the parameter estimation error is negligible. In this simulation design, we generate data according to the following two data generating processes (DGPs):

\[ y_t = \beta_1 y_{t-1}(1 + \exp[-c_1(y_{t-1} - \pi_1)])^{-1} + v_{1t} \] (14)

and

\[ y_t = \beta_2 y_{t-1}(1 + \exp[-c_2(y_{t-1} - \pi_2)])^{-1} + v_{2t} \] (15)

For the DGP in 14, we successively consider the following values for \( b_1 \), the parameter that indexes the magnitude of \( \beta_1 \): \( b_1 = [4, 6, 8, 10, 12] \). In finite-samples, \( b_1 = \beta_1 \sqrt{R} \). Thus, with an in-sample size of \( R = 500 \), the \( b_1 \) values correspond to true values of \( \beta_{0,1} = [0.18, 0.26, 0.35, 0.45, 0.53] \). In the DGP in 15, \( b_2 \) is fixed to 12 which implies that \( \beta_{0,2} = 0.53 \), a value for which, according to the simulation results of the previous section, the model is already strongly identified. The true values of the other parameters are set to \( c_{0,1} = c_{0,2} = 10 \) and \( \pi_{0,1} = \pi_{0,2} = -1.5 \) and the error terms \( v_{1t} \) and \( v_{2t} \) are drawn from a \( N(0,1) \) distribution. To simplify computations, in 14 and 15 we only consider the nonlinear part of the LSTAR(1) model.

Note that in this simulation exercise, we are not interested in results given by very small \( b_1 \) values, such as \( b_1 = 0 \) or \( b_1 = 2 \), because in that case \( \beta_1 \) is small too, and one would not be interested in forecasting with such small values of \( \beta_1 \).

The two competing models considered for forecast evaluation are the model in 14, which can have different strengths of identification depending on \( b_1 \), and the model in 15, which is always identified. When comparing the predictive ability of 14 and 15, based for example on the mean square forecast errors (MSE), the null of equal predictive ability should be satisfied. The number of simulation repetitions in the computation of the out-of-sample predictive ability statistic is 500.

Figure 5 provides the finite-sample densities of the out-of-sample test statistic, under the null of equal predictive ability, for different \( b_1 \) values in 14. For smaller \( b_1 \) values, the
model 1 (equation 14) is not (strongly) identified and the parameters are not correctly estimated. Thus, the forecast errors generated from this model are larger than the errors obtained from model 2 (equation 15). As a consequence, the null distribution of the out-of-sample predictive evaluation test is biased to the right and it is not well approximated by the standard normal distribution. A standard normal critical value could lead to misleading inference. For example, when $b_1 = 4$, the 0.95 quantile of the distribution of the test statistic is 3.11, whereas for $b_1 = 12$, the 0.95 quantile is roughly 2.

![Figure 5: Finite sample densities of the Diebold-Mariano test statistic](image)

Figure 6 reports the finite sample coverage probabilities of the nominal 95% standard confidence intervals, constructed based on the standard normal critical value, for different $b_1$ values for model 1\(^\text{13}\). The confidence intervals are constructed by inverting the test statistic. For a number of 500 simulation repetitions, we collect all values of the statistic that in absolute value are less than 1.96, for each discrete value of $b_1$. Then, the coverage probability is obtained as the proportion of instances in which the value of the statistic is contained in the standard confidence interval. The smaller is the $b_1$ value (i.e. the lower is the identification strength), the lower is the coverage probability of the standard confidence interval. The smallest finite-sample coverage probabilities are around 0.81 for $b_1 = 4$. As $b_1$ increases, the coverage probabilities are progressively approaching 0.95.

\(^{13}\)The discrete values of $b_1$ for which computations are made run from 4 to 12 with a grid of 0.2.
4 Asymptotic Results under Semi-Strong Identification

We are interested in the asymptotic distribution of the statistic $S_P$, when one or both models belong to the class in 1, and thus can suffer from identification deficiencies, in settings where the parameter estimation error is non-negligible.

When $\pi_1$ and $\pi_2$ are non-identified or weakly identified, it is known from the in-sample asymptotic results of AC2012 that their extremum estimators are inconsistent. In consequence, the framework in West (1996) cannot be applied in order to derive the asymptotic distribution of $S_P$.\footnote{More precisely, Theorem 4.1 in West (1996) cannot be applied.} Finite-sample distributions of out-of-sample tests can still be analysed in this case through simulation, as in Section 3.\footnote{Note that a STAR model with $\pi$ non-identified or weakly identified with very small beta estimates is not of interest in a prediction evaluation framework. This is not however the case for the ARMA(1,1) model.} When $\pi_1$ and $\pi_2$ are strongly identified, the asymptotic distribution of $S_P$ is of course standard.

In this section, we derive the asymptotic distribution of $S_P$ when $\pi_1$ and $\pi_2$ are semi-strongly identified\footnote{When $\pi$ is semi-strongly identified, its extremum estimator is consistent, see Theorem 3.2 in Andrews and Cheng (2012).} and show that the asymptotic distribution of the out-of-sample test
under the null of equal predictive ability is standard, even though we allow \( \pi_1 \) and \( \pi_2 \) to be only semi-strongly identified. This is different from West (1996) where all parameters are assumed to be strongly identified. Semi-strong identification does not affect the asymptotic distribution of \( S_P \), even though in this case, \( \hat{\pi}_{1R} \) and \( \hat{\pi}_{2R} \) have a slower rate of convergence, \( \sqrt{R}\|\beta_{1R}\| \ll \sqrt{R} \) and \( \sqrt{R}\|\beta_{2R}\| \ll \sqrt{R} \), than it is standard.\(^{17}\) This is in conformity with the asymptotic results in AC2012 and Cheng (2015), regarding the asymptotic distributions of in-sample statistics such as \( t \), Wald, QLR, which are shown to have a standard distribution under the null, once \( \pi \) is at least in the semi-strong identification category. These statistics are shown to have non-standard distributions only when \( \pi \) is non-identified or weakly identified.

We derive our results for the case when both models belong to the class of models under consideration and both are semi-strongly identified. When only one of the competing models belongs to this class, the same result obtains and the derivations corresponding to the standard model (the one without identification deficiencies) are as in West (1996).

The following assumptions are required to derive our result:

**Assumption 1** (Parameter space):
(i) The true parameter space corresponding to model 1, denoted \( \Theta_1^* \), lies in the interior of the optimization parameter space \( \Theta_1 \).
(ii) The true parameter space corresponding to model 2, denoted \( \Theta_2^* \), lies in the interior of the optimization parameter space \( \Theta_2 \).

Assumption 1 ensures that the true value of the parameter cannot lie on the boundary of the optimization parameter space. This way, we exclude possible boundary effects and the focus is solely on the effects of identification deficiencies.

**Assumption 2** (Criterion function):
(i) If \( \beta_1 = 0 \), \( Q_R(\theta_1) \) does not depend on \( \pi_1 \), \( \forall \theta_1 = (\beta_1, \zeta_1, \pi_1) = (0, \zeta_1, \pi_1) \in \Theta_1, \forall R \geq 1. \)
(ii) If \( \beta_2 = 0 \), \( Q_R(\theta_2) \) does not depend on \( \pi_2 \), \( \forall \theta_2 = (\beta_2, \zeta_2, \pi_2) = (0, \zeta_2, \pi_2) \in \Theta_2, \forall R \geq 1. \)

Assumption 2 is a key assumption in the paper. It is the main characteristic of the class of models under consideration. Under this assumption, the criterion function is flat with respect to \( \pi_1 \) (or \( \pi_2 \)) when \( \beta_1 \) (or \( \beta_2 \)) is close to zero and the second derivative matrix

\(^{17}\)See Theorem 3.2 in Andrews and Cheng (2012).
of the criterion function is singular or near singular.

The next three assumptions, are similar to assumptions D1, D2, and D3 in AC2012 and concern the quadratic expansion that the criterion function must satisfy when $\pi_1$ and $\pi_2$ are semi-strongly identified, and, the behavior of the first and second derivatives of the criterion function.

**Assumption 3** (Quadratic expansion of the criterion function):

(i) The sample criterion function $Q_R(\theta_1)$ has a quadratic expansion in $\theta_1$ around the true value $\theta_1^R$, i.e.

$$Q_R(\theta_1) = Q_R(\theta_1^R) + DQ_R(\theta_1^R)'(\theta_1 - \theta_1^R) + \frac{1}{2}(\theta_1 - \theta_1^R)'D^2Q_R(\theta_1^R)(\theta_1 - \theta_1^R) + R_R(\theta_1)$$

where $DQ_R(\theta_1^R) \in \mathbb{R}^{d_1}$ is a stochastic generalized first derivative vector, $D^2Q_R(\theta_2^R) \in \mathbb{R}^{d_1 \times d_1}$ is a generalized second partial derivative matrix and $R_R(\theta_1)$ is a remainder term.

(ii) The sample criterion function $Q_R(\theta_2)$ has a quadratic expansion in $\theta_2$ around the true value $\theta_2^R$, i.e.

$$Q_R(\theta_2) = Q_R(\theta_2^R) + DQ_R(\theta_2^R)'(\theta_2 - \theta_2^R) + \frac{1}{2}(\theta_2 - \theta_2^R)'D^2Q_R(\theta_2^R)(\theta_2 - \theta_2^R) + R_R(\theta_2)$$

where $DQ_R(\theta_2^R) \in \mathbb{R}^{d_2}$ is a stochastic generalized first derivative vector, $D^2Q_R(\theta_2^R) \in \mathbb{R}^{d_2 \times d_2}$ is a generalized second partial derivative matrix and $R_R(\theta_2)$ is a remainder term.

The next assumption requires good behavior of the rescaled generalized second derivative of $Q_R(\theta_1)$ and $Q_R(\theta_2)$, to eliminate its singularity when $\beta_1 \to 0$ and $\beta_2 \to 0$, which occurs when $\pi_1$ and $\pi_2$ are semi-strongly identified.

**Assumption 4** (Behavior of the generalized second derivative):

(i) $J_{1R} = B^{-1}(\beta_1^R)D^2Q_R(\theta_1^R)B^{-1}(\beta_1^R) \xrightarrow{p} J_1(\gamma_{0,1}) \in \mathbb{R}^{d_1 \times d_1}$, where $J_1(\gamma_{0,1})$ is nonsingular and symmetric.

(ii) $J_{2R} = B^{-1}(\beta_2^R)D^2Q_R(\theta_2^R)B^{-1}(\beta_2^R) \xrightarrow{p} J_2(\gamma_{0,2}) \in \mathbb{R}^{d_2 \times d_2}$, where $J_2(\gamma_{0,2})$ is nonsingular and symmetric.

The matrices $B(\beta_1)$ and $B(\beta_2)$, defined in the appendix, are used to normalize $D^2Q_R(\theta_1^R)$ and $D^2Q_R(\theta_2^R)$ so that they are nonsingular asymptotically.

The following assumption requires the rescaled generalized first derivative to satisfy a
CLT.

Assumption 5 (Behavior of the generalized first derivative):
(i) $R^{1/2}B^{-1}(\beta_1)DQ_R(\theta_1) \xrightarrow{d} \mathcal{G}^*(\gamma_{0.1}) \sim N(0_{d_{\theta_1}}, V(\gamma_{0.1}))$ for a $d_{\theta_1} \times d_{\theta_1}$ symmetric positive definite matrix $V(\gamma_{0.1})$.
(ii) $R^{1/2}B^{-1}(\beta_2)DQ_R(\theta_2) \xrightarrow{d} \mathcal{G}^*(\gamma_{0.2}) \sim N(0_{d_{\theta_2}}, V(\gamma_{0.2}))$ for a $d_{\theta_2} \times d_{\theta_2}$ symmetric positive definite matrix $V(\gamma_{0.2})$.

The next two assumptions are standard as in West (1996). They require measurability and differentiability of the loss function $g$, impose conditions on the moments of $g$ and impose stationarity and mixing conditions.

Assumption 6 (i) In some neighborhood $N_1$ around $\theta_1$, with probability one, $g(u_{t+h})$ is measurable and twice continuously differentiable with respect to $\theta_1$.
(ii) In some neighborhood $N_2$ around $\theta_2$, with probability one, $g(\epsilon_{t+h})$ is measurable and twice continuously differentiable with respect to $\theta_2$.
(iii) For all $t$, there exists a constant $D < \infty$, such that $\sup_{\theta_1 \in N_1} \| \nabla^2_{\theta_1} g(u_{t+h}) \| \leq n_{1t}$, for a measurable function $n_{1t}$ for which $E(n_{1t}) < D$.
(iv) For all $t$, there exists a constant $D < \infty$, such that $\sup_{\theta_2 \in N_2} \| \nabla^2_{\theta_2} g(\epsilon_{t+h}) \| \leq n_{2t}$, for a measurable function $n_{2t}$ for which $E(n_{2t}) < D$.

Assumption 7 (i) For some $d > 1$, $\sup_t E[\| \nabla_{\theta_1} g(u_{t+h}), g(u_{t+h}), \nabla_{\theta_2} g(\epsilon_{t+h}), g(\epsilon_{t+h}) \|^2] < \infty$, where $\| \cdot \|$ is the Euclidean norm.
(ii) $\| (\nabla_{\theta_1} g(u_{t+h}) - E(\nabla_{\theta_1} g(u_{t+h})), (g(u_{t+h}) - E(g(u_{t+h})), (\nabla_{\theta_2} g(\epsilon_{t+h}) - E(\nabla_{\theta_2} g(\epsilon_{t+h})), (g(\epsilon_{t+h}) - E(g(\epsilon_{t+h})), \| \|^2$ is strong mixing with mixing coefficients of size $-3d/(d - 1)$.
(iii) $\| \nabla_{\theta_1} g(u_{t+h}), g(u_{t+h}), \nabla_{\theta_2} g(\epsilon_{t+h}), g(\epsilon_{t+h}) \|^2$ is covariance stationary.
(iv) Let $\Gamma_{t+h} = g(u_{t+h}) - g(\epsilon_{t+h})$ and let $V_\epsilon = \sum_{j=-\infty}^{+\infty} E(\Gamma_{t+h} - E(\Gamma_{t+h}))(\Gamma_{t+h-j} - E(\Gamma_{t+h-j}))$, then $V_\epsilon$ is positive definite.

The last assumption is standard in the predictive ability testing literature and concerns the relative rate of increase of $T$, $R$ and $P$.

Assumption 8 $P, R \to \infty$ as $T \to \infty$ and $\lim_{T \to \infty} \frac{P}{R} = \tilde{\pi}$ with $0 \leq \tilde{\pi} < \infty$.

The following result provides the asymptotic distribution of $S_P$ for the case in which $\pi_1$ and $\pi_2$ are semi-strongly identified. The proof of the theorem is given in Appendix 1.
Theorem 1. Let Assumptions 1-8 hold. Under $H_0$,

$$S_P \xrightarrow{d} N(0, \Omega)$$

where

$$\Omega = V_\epsilon + \lambda_1 D_{\theta_1} V_{W_1} D_{\theta_2} + \lambda_1 D_{\theta_1} V_{W_2} D_{\theta_2} + 2\lambda_2 D_{\theta_1} C_{\epsilon W_1} - 2\lambda_2 D_{\theta_2} C_{\epsilon W_2} - 2\lambda_1 D_{\theta_1} C_{W_1 W_2} D_{\theta_2}$$

with:

$$V_\epsilon = \sum_{j=+\infty}^{j=-\infty} E[(g(u_{t+h}) - g(\epsilon_{t+h})) - g(u_{t+h}) - g(\epsilon_{t+h}))]$$

$$V_{W_1} = -J^{-1}_1(\gamma_{0,1})B^{-2}(\beta_{1R})\left(\sum_{j=+\infty}^{j=-\infty} m_{1,t,m_{1,t+j}}\right)B^{-2}(\beta_{2R})(-J^{-1}_1(\gamma_{0,1}))$$

$$V_{W_2} = -J^{-1}_2(\gamma_{0,2})B^{-2}(\beta_{2R})\left(\sum_{j=+\infty}^{j=-\infty} m_{2,t,m_{2,t+j}}\right)B^{-2}(\beta_{2R})(-J^{-1}_2(\gamma_{0,2}))$$

$$C_{\epsilon W_1} = \left(\sum_{j=+\infty}^{j=-\infty} E[(g(u_{t+h}) - g(\epsilon_{t+h}))m_{1,t+j}]\right)B^{-2}(\beta_{1R})(-J^{-1}_1(\gamma_{0,1}))$$

$$C_{\epsilon W_2} = \left(\sum_{j=+\infty}^{j=-\infty} E[(g(u_{t+h}) - g(\epsilon_{t+h}))m_{2,t+j}]\right)B^{-2}(\beta_{2R})(-J^{-1}_2(\gamma_{0,2}))$$

$$C_{W_1 W_2} = -J^{-1}_1(\gamma_{0,1})B^{-2}(\beta_{1R})\left(\sum_{j=+\infty}^{j=-\infty} m_{1,t,m_{2,t+j}}\right)B^{-2}(\beta_{2R})(-J^{-1}_2(\gamma_{0,2}))$$
and $D_{\theta_1} = E(\nabla_{\theta_1} g(u_{t+h}))$, $D_{\theta_2} = E(\nabla_{\theta_2} g(\epsilon_{t+h}))$.

The parameters $\lambda_1$ and $\lambda_2$ are defined as follows. In a rolling estimation scheme, we have that $\lambda_1 = (1 - \frac{1}{2\pi})$ and $\lambda_2 = (1 - \frac{1}{2\pi})$, in the case where $1 < \tilde{\pi} < \infty$ and, $\lambda_1 = (\tilde{\pi} - \frac{2\pi^2}{3})$ and $\lambda_2 = \frac{\tilde{\pi}}{2}$, in the case where $\tilde{\pi} \leq 1$. If we consider a recursive estimation scheme, $\lambda_1 = 2(1 - \frac{1}{\tilde{\pi}} \ln(1 + \tilde{\pi}))$ and $\lambda_2 = 1 - \frac{1}{\tilde{\pi}} \ln(1 + \tilde{\pi})$. For a fixed window estimation scheme, the parameters $\lambda_1$ and $\lambda_2$ are $\lambda_1 = \tilde{\pi}$, $\lambda_2 = 0$.

Theorem 1 shows that even though $\pi_1$ and $\pi_2$ might be only semi-strongly identified, the asymptotic distribution of $S_P$, based on the class of models defined in (1), is the same as in West (1996). Note that the results in West (1996) are derived assuming OLS estimation, while here we assume a maximum likelihood estimation framework.

5 Robust Inference

As shown in the numerical finite-sample results (Section 3), a standard strong-identification critical value can lead to misleading inference, when $\pi_1$ (and/or $\pi_2$) are not (strongly) identified. In this section, we construct robust critical values that take into account the possible loss of (strong) identification. A robust critical value in our testing framework is larger than the standard strong identification critical value. Robust confidence intervals can also be constructed by inverting the test statistic using the robust critical value. Here the focus is on cases in which the parameter estimation error is negligible. Robust critical values in non-negligible parameter estimation settings can be obtained by bootstrap inference. This is considered in future research.

5.1 Data Dependent Critical Values

The first step in the construction of robust critical values is to employ an identification selection procedure (ICS) that uses the data to determine whether $\|b\| < \infty$ (i.e. non-identification/weak identification) or $\|b\| = \infty$ (i.e. semi-strong/strong identification).\textsuperscript{18}

The identification selection procedure distinguishes between weak and semi-strong

\textsuperscript{18}This identification selection procedure is also employed in Andrews and Cheng (2012) and Cheng (2015) and is related to methods proposed in Andrews (1999) and Andrews and Soares (2010).
identification of $\pi$ based on the following statistic:\footnote{For ease of notation, the indices 1 and 2 corresponding to model 1 and 2 are omitted in this subsection.}

$$ICS_R = (R \beta_R' \hat{\Sigma}_{\beta \beta,R}^{-1} \beta_R / d_\beta)^{1/2}$$

(18)

where $\hat{\Sigma}_{\beta \beta,R}$ is the the upper left $d_\beta \times d_\beta$ block of $\hat{\Sigma}_R$ - the estimator of the variance-covariance matrix of $\hat{\theta}_R$, and $R$ is the in-sample size\footnote{In order to decrease computation times, when computing the ICS statistic, a fixed estimation scheme is used.}. Note that when $\beta_R = O(1/\sqrt{R})$, (i.e. when $\pi$ is weakly identified), $ICS_R$ is $O_p(1)$.

We select the weak identification category if

$$ICS_R \leq k_R$$

(19)

and the semi-strong identification category, otherwise.\footnote{As $ICS_R$ is $O_p(1)$ when $\pi$ is weakly identified, one consistently selects the weak identification category, provided $k_R$ diverges to infinity.} The sequence of constants $k_R : R \geq 1$ are tuning parameters such that:

$$i) \ k_R \to \infty \quad \text{and} \quad ii) \ k_R/R^{1/2} \to 0.$$ 

(20)

For example, $k_R$ can be taken $k_R = (\ln R)^{1/2}$, which is analogous to the Bayesian information criterion (BIC) penalty term, to satisfy i) and ii).

To understand why the second condition on $k_R$, (ii), is also needed, consider a strong identification case when $\beta_R$ is bounded away from 0, say $\beta_R = 2$. In this case, $\hat{\beta}_R$ converges to 2 in probability and $ICS_R$ diverges to infinity at rate $\sqrt{R}$. To ensure $ICS_R$ is larger than $k_R$, we need $k_R$ to diverge at a rate slower than $\sqrt{R}$, thus the second condition $k_R/R^{1/2} \to 0$.

Using the identification selection procedure described above, a robust critical value, at nominal level $1 - \alpha$, $\hat{c}_{R,1-\alpha}$ is defined as:

$$\hat{c}_{R,1-\alpha} = \begin{cases} 
  c_{1-\alpha}^{LF} & \text{if } ICS_R \leq k_R \\
  c_{1-\alpha}^\infty & \text{if } ICS_R > k_R
\end{cases}$$

(21)

where $c_{1-\alpha}^\infty$ is the standard strong identification critical value, and $c_{1-\alpha}^{LF}$ is the least favor-
able (LF) critical value, that is large enough for all identification categories:

\[ c_{1-\alpha}^{LF} = \max \left\{ \sup_{l \in L} c_{1-\alpha}(l), \ c_1^{\infty} \right\}. \]  \hspace{1cm} (22)

In (22), \( c_{1-\alpha}(l) \) is the \( 1 - \alpha \) quantile of the statistic under weak/non-identification, and \( l = (b, \gamma) \in L \), with \( L = \{ l = (b, \gamma) : \|b\| < \infty, \gamma \in \Gamma \} \). Note that one could directly use, \( c_{1-\alpha}^{LF} \), the least favorable critical value as the robust critical value, without the identification selection procedure, but the associated confidence interval is typically overly long, and not as informative in the strong identification case. The ICS procedure improves on the LF critical value.

The robust critical value, \( \hat{c}_{R,1-\alpha} \), can be further improved by employing null-imposed and/or plug-in versions of it. A null imposed critical value exploits the knowledge of the null hypothesis value of a restriction, \( r(\theta) \), implied by the test, such that the null is \( H_0 : r(\theta) = v \) for \( v \in r(\Theta) \). For example, in a context where the null hypothesis specifies a value \( \pi^* \) for \( \pi \), then the supremum in (22) needs to be taken only over the \( l \) values for which \( \pi = \pi^* \). In this case, \( L \) is replaced by \( L(v) = \{ l = (b, \gamma) \in L : \|b\| < \infty, \ r(\theta) = v \} \) in (22).

The plug-in version of \( \hat{c}_{R,1-\alpha} \) can be employed when part of \( \gamma = (\zeta, \beta, \pi, \phi) \) is unknown under \( H_0 \), but can be consistently estimated. This critical value replaces elements of \( \gamma \) with consistent estimators in the formula in (22), and the supremum over the set \( L \) can be reduced to a supremum over \( \hat{L}_R \) (or \( L(v) \cap \hat{L}_R \)), where \( \hat{L}_R = \{ l = (b, \gamma) \in L : \gamma = (\hat{\zeta}_R, \beta, \pi, \phi) \} \), as \( \zeta \) is consistently estimated by \( \hat{\zeta}_R \).

5.2 Simulation: Coverage Probabilities based on Robust Critical Values

Consider again the simulation design from Section 3.2. We apply the methods described above to construct robust critical values and thus obtain robust tests. We consider both the least favorable critical value \( c_{1-\alpha}^{LF} \) as well as the data dependent critical value \( \hat{c}_{R,1-\alpha} \).

The least favorable critical value \( c_{1-\alpha}^{LF} \) is obtained as follows. Fix \( \pi_{0,1} \) in 14 successively

---

\footnote{The null imposed critical value is not particularly useful in our setup, as the null involves the loss differential, see equation (9), and does not impose a particular value for the parameters.}

\footnote{The plug in version of \( \hat{c}_{R,1-\alpha} \) is more suitable to our testing framework.}
to one of the values \( \pi_{0,1} = [-3.5, -3, -2.5, -2, -1.5] \). For each of these \( \pi_{0,1} \), run the \( b_1 \) values, from 4 to 10 with a grid of 0.2, obtain the distribution of the test statistic and compute the 0.95 quantile. The maximum of these quantiles over all values of \( b_1 \) and \( \pi_{0,1} \) gives the least favorable critical value. Note that in this simulation exercise, a value of \( \zeta = (\zeta_1, \zeta_2)' \) is not required in order to obtain \( c_{1-a}^{LF} \). Because \( \zeta \) can be consistently estimated, in applications where \( \zeta \) appears, one could plug-in the estimated value of \( \zeta \) in place of \( \zeta_0 \).

Figure 7 reports the finite-sample coverage probabilities of confidence intervals constructed based on the least favorable critical value as a function of \( b_1 \). The least favorable critical value is obtained for \( b_1 = 4 \) and \( \pi_{0,1} = -2 \), and has a value of 3.15. The robust confidence interval is obtained by inverting the test statistic using the least favorable critical value. The coverage probabilities are closer to 0.95 for smaller values of \( b_1 \). For larger \( b_1 \) values, the coverage probabilities are 1, as the robust confidence interval based on the least favorable critical value is overly long when model 1 is strongly identified.

The data dependent robust critical value \( c_{1-a}^{LF} \) that employs an identification category selection procedure does not lead to coverage probabilities of 1, once \( \pi_1 \) is at least semi-strongly identified because the robust critical value switches to the standard critical value.

Figure 8 reports the coverage probabilities computed based on the data dependent robust critical value. For \( R = 500 \), \( k_R = (\ln R)^{1/2} = 2.49 \). This value is compared with the value of the \( ICS_R \) statistic obtained over the grid values of \( b_1 \). The identification category selection statistic becomes larger than \( k_R \) at \( b_1 = 11.2 \). Using the data dependent robust critical value the coverage probabilities are closer to 0.95 for larger values of \( b_1 \), where \( \pi_1 \) is at least semi strongly identified. In this procedure, the transition from the least favorable critical value to the standard critical value is not continuous.

\(^{24}\text{These are values from the true parameter space of } \pi.\)
Figure 7: Coverage probabilities of robust CIs with least favorable critical value

Figure 8: Coverage probabilities of robust CIs with data dependent critical values
6 Conclusion

The paper analyzes the predictive accuracy evaluation of models affected by identification deficiencies. The focus is on a class of models that are strongly identified in some part of the parameter space but non-identified or weakly identified in another part of the parameter space. The potential source of strong identification loss is an explicit part of the model.

Numerical results show that when identification is lost, the finite-sample distributions of estimators are far from the normal distribution, they can be bimodal or uniform, for example. As a result, the forecast errors obtained under identification loss are larger than the errors obtained under strong identification and, in situations in which the parameter estimation is negligible, the null distribution of out-of-sample predictive ability tests is not well approximated by the standard normal distribution. Thus, out-of-sample tests and confidence intervals obtained by inverting tests that employ the standard normal critical value have size distortions and can lead to misleading inference.

We propose two methods to construct robust critical values for this problem. The first is based on the least favorable possible distribution of the test statistic under the null, whereas the second is data dependent and uses an identification category selection procedure to determine the identification strength.

In settings where the parameter estimation error is non-negligible, the paper shows that the asymptotic distribution of the out-of-sample predictive ability test in West (1996) is standard, even when one allows for the model(s) to be only semi-strongly identified.
References


_ , _ , and Patrik Guggenberger, “Generic results for establishing the asymptotic size of confidence sets and tests,” 2011.


Appendix 1:

Proof of Theorem 1:

Taking a Taylor expansion of the first term of $S_P$ around the probability limit $\theta_{1R}$, we have that:

$$\frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} g(u_{t+h}) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} g(u_{t+h}) + \frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} \nabla_{\theta_1} g(u_{t+h})'(\hat{\theta}_{1R,t} - \theta_{1R}) + o_p(1) \quad (23)$$

Denote

$$Z_{1R} = -R^{1/2}J_{1R}^{-1}B^{-1}(\beta_{1R})DQ_R(\theta_1) \quad (24)$$

and

$$\Delta_{1R} = R^{1/2}B(\beta_{1R})(\hat{\theta}_{1R} - \theta_{1R}) \quad (25)$$

The matrix $B(\beta_1)$ in (24) and (25) is defined as $B(\beta_1) = \begin{bmatrix} I_{d_{\psi_1}} & 0_{d_{\psi_1} \times d_{\pi_1}} \\ 0_{d_{\psi_1} \times d_{\pi_1}} & \iota(\beta_1)I_{d_{\pi_1}} \end{bmatrix} \in \mathbb{R}^{d_{\psi_1} \times d_{\pi_1}}$, with $\iota(\beta_1) = \beta_1$ if $\beta_1$ is a scalar and $\iota(\beta_1) = \|\beta_1\|$ if $\beta_1$ is a vector, and is used to normalize $D^2Q_R(\theta_{1R})$, so that it is nonsingular asymptotically, when $\beta_{1R} \to 0$ (which is when $\pi_1$ is semi-strongly identified).

From the proof of Theorem 3.2 in *AC2012*, that provides the asymptotic distribution of the in-sample estimator of $\theta$, in the semi-strong identification case, we have that:$^{25}$

$$\Delta_{1R} = Z_{1R} + o_p(1) \quad (26)$$

Combining (24), (25) and (26) yields:

$$(\hat{\theta}_{1R} - \theta_{1R}) = -J_{1R}^{-1}B^{-2}(\beta_{1R})DQ_R(\theta_1) + o_p(1) \quad (27)$$

$^{25}$In the equations below, the index $R$ is the size of the rolling sample.
Now, in a rolling estimation setup, we can write that for each rolling window \( R \leq t \leq T \):

\[
(\hat{\theta}_{1,R,t} - \theta_{1R}) = -J_{1R}^{-1}B^{-2}(\beta_{1R})D^{(t)}Q_{R}(\theta_1) + o_p(1) \quad (28)
\]

Substituting (28) in (23) and using Assumption 4, we have:

\[
\frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} g(\hat{u}_{t+h}) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} g(u_{t+h}) + D'_{\theta_{1}}(-J_{1}^{-1}(\gamma_{0,1}))\frac{1}{\sqrt{R}} \sum_{t=R}^{T-h} B^{-2}(\beta_{1R})D^{(t)}Q_{R}(\theta_1) \\
+ O_p\left(\frac{1}{\sqrt{P}}\right) + o_p(1) \quad (29)
\]

as \( \frac{1}{P} \sum_{t=R+1}^{T-h} \nabla_{\theta_1} g(u_{t+h})' - D_{\theta_1} = O_p\left(\frac{1}{\sqrt{P}}\right) \) with \( D_{\theta_1} = E(\nabla_{\theta_1} g(u_{t+h})) \).

In a rolling estimation case, the score \( D^{(t)}Q_{R}(\theta_1) \) is of the form \( \frac{1}{R} \sum_{j=t-R+1}^{t} m_{1,j} \), where \( m_{1,j} \in \mathbb{R}^{d_{\theta_1}} \) is a function of \( \theta_1 \) and the observables. Thus, we can write:

\[
\frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} g(\hat{u}_{t+h}) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} g(u_{t+h}) + D'_{\theta_{1}}(-J_{1}^{-1}(\gamma_{0,1}))B^{-2}(\beta_{1R}) \sqrt{\frac{1}{P}} \sum_{t=R}^{T-h} \sum_{j=t-R+1}^{t} m_{1,j} \\
+ O_p\left(\frac{1}{\sqrt{P}}\right) + o_p(1) \quad (30)
\]

Considering now the second term of the statistic \( S_P \). Analogous to the derivations above we can write that:

\[
\frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} g(\hat{\epsilon}_{t+h}) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} g(\epsilon_{t+h}) + D'_{\theta_{2}}(-J_{2}^{-1}(\gamma_{0,2}))B^{-2}(\beta_{2R}) \sqrt{\frac{1}{P}} \sum_{t=R}^{T-h} \sum_{j=t-R+1}^{t} m_{2,j} \\
+ O_p\left(\frac{1}{\sqrt{P}}\right) + o_p(1) \quad (31)
\]

Equation (30) and (31) give the expression of \( S_P \),

\[\text{Note that, in a rolling estimation scheme, for each rolling window indexed by } t, \hat{\theta}_{1,R,t} \text{ is computed based on the score, } D^{(t)}Q_{R}(\theta_1) \text{, indexed by } t.\]
\[
S_p = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} (g(u_{t+h}) - g(\epsilon_{t+h})) \\
+ D'_\theta_1 (-J^{-1}_1(\gamma_{0,1})) B^{-2}(\beta_1 R) \frac{\sqrt{P}}{R} \sum_{t=R}^{T-h} \sum_{j=t-R+1}^{t} m_{1,j} \\
- D'_\theta_2 (-J^{-1}_2(\gamma_{0,2})) B^{-2}(\beta_2 R) \frac{\sqrt{P}}{R} \sum_{t=R}^{T-h} \sum_{j=t-R+1}^{t} m_{2,j} \\
+ O_p\left(\frac{1}{\sqrt{P}}\right) + o_p(1)
\]

(32)

Now, as all estimators are consistent when \(\pi_1\) and \(\pi_2\) are semi-strongly identified, from Theorem 4.1 in West (1996), we have that under the null of \(H_0\): \(E(g(u_{t+h}) - g(\epsilon_{t+h})) = 0\),

\[
\frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} (g(\hat{u}_{t+h}) - g(\hat{\epsilon}_{t+h})) \overset{d}{\to} N(0, \Omega)
\]

where:

\[
\Omega = V + \lambda_1 D'_\theta_1 V W_1 D\theta_1 + \lambda_1 D'_\theta_1 V W_2 D\theta_2 + 2\lambda_2 D\theta_1 C \epsilon W_1 - 2\lambda_2 D\theta_2 C \epsilon W_2 - 2\lambda_1 D'_\theta_1 C W_1 W_2 D\theta_2
\]

and

\[
V_e = \sum_{j=-\infty}^{j=+\infty} E[(g(u_{t+h}) - g(\epsilon_{t+h})) - E(g(u_{t+h}) - g(\epsilon_{t+h}))] \\
\times (g(u_{t+h+j}) - g(\epsilon_{t+h+j})) - E(g(u_{t+h}) - g(\epsilon_{t+h}))]
\]

\[
V_{W_1} = -J^{-1}_1(\gamma_{0,1}) B^{-2}(\beta_1 R) \left(\sum_{j=-\infty}^{j=+\infty} m_{1,t} m'_{1,t+j}\right) B^{-2}(\beta_1 R)(-J^{-1}_1(\gamma_{0,1}))
\]

\[
V_{W_2} = -J^{-1}_2(\gamma_{0,2}) B^{-2}(\beta_2 R) \left(\sum_{j=-\infty}^{j=+\infty} m_{2,t} m'_{2,t+j}\right) B^{-2}(\beta_2 R)(-J^{-1}_2(\gamma_{0,2}))
\]

35
\[
C_{\epsilon W_1} = \left( \sum_{j=-\infty}^{j=+\infty} E[(g(u_{t+h}) - g(\epsilon_{t+h}))m_{1,t+j}'] \right) B^{-2}(\beta_{1R})(-J^{-1}_1(\gamma_{0,1}))
\]

\[
C_{\epsilon W_2} = \left( \sum_{j=-\infty}^{j=+\infty} E[(g(u_{t+h}) - g(\epsilon_{t+h}))m_{2,t+j}'] \right) B^{-2}(\beta_{2R})(-J^{-1}_2(\gamma_{0,2}))
\]

\[
C_{W_1 W_2} = -J^{-1}_1(\gamma_{0,1}) B^{-2}(\beta_{1R}) \left( \sum_{j=-\infty}^{j=+\infty} m_{1,t} m_{2,t+j}' \right) B^{-2}(\beta_{2R})(-J^{-1}_2(\gamma_{0,2}))
\]
Appendix 2: Numerical Results

Figure 9: Finite sample densities of the estimator of $\zeta_1$ when $\pi_0 = -3$. 

Figure 10: Finite sample densities of the estimator of $\zeta_2$ when $\pi_0 = -3$. 

37
Figure 11: Finite sample densities of the estimator of $\beta$ when $\pi_0 = -3$.

Figure 12: Finite sample densities of the estimator of $\pi$ when $\pi_0 = -3$. 