Nonparametric Tests for Superior Predictive Ability

Thierry Post*, Valerio Potì†, Selcuk Karabati‡& Stelios Arvanitis§

June 12, 2019

Abstract

A nonparametric method for comparing multiple forecast models is developed and implemented. The hypothesis of Optimal Predictive Ability generalizes the Superior Predictive Ability hypothesis from a single given loss function to an entire class of loss functions. Distinction is drawn between General Loss functions, Convex Loss functions and Symmetric Convex Loss functions. The Optimal Predictive Ability hypothesis is formulated in terms of moment inequality conditions. The empirical moment conditions are reduced to an exact and finite system of linear inequalities based on piecewise-linear loss functions. The hypothesis can be tested in a statistically consistent way using a blockwise Empirical Likelihood Ratio test statistic. A computationally feasible test procedure computes the test statistic using Convex Optimization methods and estimates conservative, data-dependent critical values using a majorizing chi-square limit distribution and a moment selection method. An empirical application to inflation forecasting reveals that a very large majority of thousands of forecast models are redundant, leaving predominantly Phillips Curve type models, when convexity and symmetry are assumed.

Keywords: Forecast Comparison, Stochastic Dominance, Empirical Likelihood, Inflation Forecasting

*Post is Professor of Finance at the Graduate School of Business of Nazarbayev University; Astana, Kazakhstan; e-mail: thierrypost@hotmail.com.

†Potì is Professor of Finance at the Michael Smurfit Graduate Business School of the University College Dublin; Dublin, Ireland; e-mail: valerio.poti@ucd.ie.

‡Karabatı is Professor of Operations Management at Koç University; 34450 Sarıyer/Istanbul, Turkey; e-mail: skarabati@ku.edu.tr.

§Arvanitis is Associate Professor at the Department of Economics of Athens University of Economics and Business; Athens; Greece; email: stelios@aueb.gr.

1 Introduction

A classic problem in forecasting is the comparison of a multitude of models based on different information sets and estimation methods. White (2000) and Hansen (2005) develop the standard framework for testing the hypothesis of Superior Predictive Ability (SPA).

Regretfully, the relative accuracy of forecast models is often not robust to plausible variation of the loss function. To obtain a robust classification, Jin, Corradi & Swanson (2017) generalize the SPA hypothesis from a single given loss function to an entire class of loss functions, using Stochastic Dominance (SD) orders. Their hypothesis of Stochastic Dominance Superiority states that a given forecast model dominates all alternative models. To test this hypothesis, they extend the Kolmogorov-Smirov type test of Linton, Maasoumi & Whang (2005) to forecast model comparison.

Unfortunately, the discriminatory power of the Superiority criterion quickly falls as the number of forecast models (M) increases and, inevitably, cases of non-dominance are introduced. In terms of mathematical order theory, the partially ordered set generally has multiple distinct 'maximal elements' (which are not dominated by any alternative) and hence no 'greatest element' (which dominates all alternatives).

The lack of discriminatory power is compounded by the minimal structure imposed on the permissible loss functions. Two classes were distinguished: General Loss functions and Convex Loss functions. These classes include a range of pathological loss functions which can obscure the results for standard loss functions.

To improve the discriminatory power, the present study uses an alternative generalization of the SPA hypothesis, which translates the criterion of SD Optimality (Fishburn (1974), Bawa, Bodurtha, Rao & Suri (1985) and Post (2017)) to forecast comparison and which is labeled here as Optimal Predictive Ability (OPA).

A given forecast is optimal if it minimizes expected loss for some permissible loss function. Non-optimal forecasts are suboptimal for all loss functions and can therefore be discarded

from the analysis. Importantly, a given forecast can be optimal without dominating alternative forecasts and it can be non-optimal without being dominated, which introduces additional power.

The SD Optimality criterion has been shown to reduce the number of choice alternatives from M to about $\sqrt{M}$ in other application areas. As a case in point, in Bawa, Bodurtha, Rao & Suri (1985), the optimal set consists of only 25 out of $M = 896$ New York Stock Exchange stocks. Anderson & Post (2018) report similar set reductions for comparing multiple income distributions.

Furthermore, a class of Symmetric Convex Loss functions is introduced. The additional assumption of symmetry improves the discriminatory power upon the analysis based on General Loss and Convex Loss functions. The Symmetric Convex Loss class includes the standard Laplacian, Gaussian and Huber loss functions but excludes many pathological loss functions. This class is shown to be closely related to standard Second-degree Stochastic Dominance (SSD; Hadar & Russell (1969), Hanoch & Levy (1969) and Rothschild & Stiglitz (1970)).

For each of the three classes of loss functions, the hypothesis of OPA is formulated using moment inequality conditions, which opens the way for using moment-based estimation and inference methods. Among these methods, Owen's (1988, 1990, 1991) Empirical Likelihood (EL) stands out as particularly promising for statistical inference about OPA.

EL and SD combine well due to a shared distribution-free assumption framework and the discrete structure of the 'implied distribution function', which facilitates numerical optimization. The complementary relation between SD and EL was previously recognized by Davidson & Duclos (2013), Davidson (2009), Post & Potì (2017), Post (2017) and Post, Karabati and Arvanitis (2018) in the context of welfare analysis, asset pricing and portfolio optimization.

In the area of forecast evaluation, the EL method has the additional advantage that it does not require information about the forecast error covariance matrix and thus avoids

problems with the estimation and manipulation of the covariance matrix when the number of evaluated models is large. Similarly, Hansen (2005, p. 367) eschews quadratic-form test statistics, to avoid these problems with the covariance matrix.

The loss function is treated as a partially identified, infinite-dimensional model parameter. For practical application, a discrete representation is obtained using piecewise-linear loss functions, in the spirit of Post (2003, Thm 2). Using this formulation, the empirical moment conditions can be be formulated as an exact and finite system of linear inequalities.

A blockwise Empirical Likelihood Ratio (ELR) test statistic is used to test the moment inequality conditions for time series data. The ELR statistic has important statistical optimality properties (Kitamura (2001); Canay (2010)). The blockwise implementation allows for a range of dynamic patterns, including common stationary ARMA, GARCH and stochastic volatility processes (Kitamura (1997)).

The optimization problem that has to be solved to compute the ELR test statistic is non-convex. A computational strategy is developed which alternates between one Convex Optimization problem for estimation the loss function given the probabilities and another Convex Optimization problem for estimating the probabilities given the loss function.

Conservative asymptotic critical values are derived using a majorizing chi-square limit distribution and moment selection methods. The resulting testing procedure is statistically consistent and asymptotically conservative.

The rest of this study is organized as follows. Section 2 illustrates the inferential framework, comparing and contrasting alternative hypothesis structures, and provides a small illustrative example. Section 3 delves further into the hypothesis structure of OPA, provides the empirical specification of moment conditions implied by the null hypothesis of OPA and illustrates the testing procedure. Section 4 derives its asymptotic properties under a stationary and absolutely regular time series framework. Section 5 provides details on a computational strategy that can be followed to carry out the test. Section 6 provides an illustrative application to exchange rates forecasting, extending the empirical section in Jin,

Corradi & Swanson (2017). Section 7 provides a larger scale application to inflation forecasting, extending Hansen (2005) study, to illustrate how our approach generalizes tests of SPA a la White (2000) and Hansen (2005) to a setting where the loss function is not parametric. Section 8 concludes. The Appendix contains formal proofs along with auxiliary results and further characterizations of the stochastic orders.

2 Theoretical Concepts

2.1 Forecast errors and loss functions

A random variable X is forecast using $M \geq 2$ distinct and given forecast models, generating point forecasts $\boldsymbol{Y} := (Y_1 \cdots Y_M)$. The forecasts could be constructed, for example, using predictive regression, analyst forecasts or market prices of securities. The forecast models could also include forecast combinations of multiple base forecasts.

One of the models is compared with the other $(M-1)$ models. The models are indexed such that the evaluated model takes the M-th position; the alternatives are collected in the set $\mathcal{I} := \{1, \cdots, M-1\}$.

Alternatives $i \in \mathcal{I}$ which are dominated or non-optimal (as defined below) are irrelevant for the analysis. It is recommended to detect and exclude such redundancies, if possible, to increase statistical power and reduce the computer time. Since the number of optimal alternatives in earlier studies was roughly $\sqrt{M}$, the number of redundancies can be substantial.

The forecast errors of the models are given by $\boldsymbol{E} := (E_1 \cdots E_M)$, $E_i := X - Y_i$, $i = 1, \cdots, M$. The joint cumulative distribution function (CDF) of the errors is denoted by $\mathcal{F} : \mathcal{X}^M \to [0,1]$, where $\mathcal{X} := [a,b]$, $-\infty < a < b < +\infty$.

Predictive ability is measured using expected loss $\mathbb{E}_{\mathcal{F}}[L(E_i)]$ based on a loss function $L : \mathcal{X} \to \mathbb{R}_+$. The relevant class of permissible loss functions is denoted by $\mathcal{L} \in \{\mathcal{L}_0, \mathcal{L}_1, \mathcal{L}_2\}$.

The class of General Loss functions, $\mathcal{L}_0$, contains all right-continuous loss functions which

achieve a minimum at $L(0) = 0$ and do not decrease as the error moves away from zero. The subset of Convex Loss functions $\mathcal{L}_1 \subset \mathcal{L}_0$ assumes also convexity: $wL(E_1) + (1-w)L(E_2) \geq L(wE_1 + (1-w)E_2)$ and continuity at the boundary points. The Symmetric Convex Loss functions $\mathcal{L}_2 \subset \mathcal{L}_1$ furthermore exhibit symmetry: $L(E) = L(|E|)$. The latter class includes the standard Laplacian, Gaussian and Huber loss functions.

The Symmetric Convex Loss class $\mathcal{L}_2$ is closely related to SSD. Specifically, $U(x) := -L(-x)$, $x \leq 0$, is an increasing and concave utility function and the minimization of $\mathbb{E}_{\mathcal{F}}[L(E)]$ is equivalent to the maximization of $\mathbb{E}_{\mathcal{F}}[U(-|E|)] = -\mathbb{E}_{\mathcal{F}}[L(|E|)]$, where U is an increasing and concave utility function. OPA of the forecast error E of a given forecast model for $\mathcal{L}_2$ thus corresponds to SSD optimality of the negative absolute forecast error $(-|E|)$ of the model.

For numerical reasons, the set of permissible functions can be reduced to piecewise-linear functions with kinks at the atoms of the empirical distribution of the evaluated forecast model. This set reduction does not affect the truth of the empirical moment condition, the value of the test statistic or the estimated critical values; see Section 3.

2.2 Stochastic Dominance

In pairwise comparisons, model $i \in \mathcal{I}$ stochastically dominates model M for loss function class $\mathcal{L}$, or $E_i \succeq_{\mathcal{L},\mathcal{F}} E_M$, if $\mathbb{E}_{\mathcal{F}}[L(E_M)] \leq \mathbb{E}_{\mathcal{F}}[L(E_i)]$ for all $L \in \mathcal{L}$; non-dominance occurs if $\mathbb{E}_{\mathcal{F}}[L(E_M)] > \mathbb{E}_{\mathcal{F}}[L(E_i)]$ for some $L \in \mathcal{L}$.

The distinction between strict and weak inequality is inconsequential for the present analysis, and $\mathbb{E}_{\mathcal{F}}[L(E_i)] \geq \mathbb{E}_{\mathcal{F}}[L(E_M)]$ can replace $\mathbb{E}_{\mathcal{F}}[L(E_i)] > \mathbb{E}_{\mathcal{F}}[L(E_M)]$ without harm. This replacement would be prohibited if the loss functions were allowed to be constant on the interior of the support of the evaluated model $\mathcal{X}_M := [a_M, b_M]$, in which case $\mathbb{E}_{\mathcal{F}}[L(E_M)] = 0$ and thus $\mathbb{E}_{\mathcal{F}}[L(E_i)] \geqslant \mathbb{E}_{\mathcal{F}}[L(E_M)]$ would become trivial.

2.3 Admissibility, Optimality and Superiority

The concept of dominance can be extended in several distinct ways to a joint analysis of all models. The three extensions also represent three distinct ways to generalize the SPA hypothesis from a given loss function to the entire class of loss functions ($\mathcal{L}$).

SD Admissibility occurs when the evaluated model is not dominated by any alternative, $\forall i \in \mathcal{I}, \exists L \in \mathcal{L} : \mathbb{E}_{\mathcal{F}}\left[L(E_i) - L(E_M)\right] > 0$. This occurrence is equivalent to the following condition:

$$\mathcal{A}(\mathcal{L}, \mathcal{F}) : \inf_{i \in \mathcal{I}} \sup_{L \in \mathcal{L}} \mathbb{E}_{\mathcal{F}}\left[L(E_i) - L(E_M)\right] > 0. \tag{1}$$

Using the terminology of mathematical order theory, an admissible model is a 'maximal element' of the partially ordered set defined by the choice set and the dominance relation.

By contrast, OPA occurs if the evaluated model minimizes expected loss for some permissible loss function, $\exists L \in \mathcal{L}, \forall i \in \mathcal{I} : \mathbb{E}_{\mathcal{F}}\left[L(E_i) - L(E_M)\right] > 0$, or equivalently:

$$\mathcal{O}(\mathcal{L}, \mathcal{F}) : \sup_{L \in \mathcal{L}} \inf_{i \in \mathcal{I}} \mathbb{E}_{\mathcal{F}}\left[L(E_i) - L(E_M)\right] > 0. \tag{2}$$

It follows from the Max-Min Inequality that admissibility is a necessary but not sufficient condition for OPA: $\mathcal{A}(\mathcal{L}, \mathcal{F}) \Leftarrow \mathcal{O}(\mathcal{L}, \mathcal{F})$. The distinction between admissibility and OPA is not trivial. The optimal set of stocks in Bawa, Bodurtha, Rao & Suri (1985) is about 30 percent smaller than the corresponding admissible set.

Again, the distinction between strict and weak inequality is inconsequential for the analysis, as the loss function is required to be increasing on $\mathcal{X}_M$. A binding inequality occurs

if the evaluated model is a non-unique minimizer due to the existence of multiple optimal models for the optimal loss function.

The nested structure $\mathcal{L}_0 \supset \mathcal{L}_1 \supset \mathcal{L}_2$ furthermore implies $\mathcal{O}(\mathcal{L}_0, \mathcal{F}) \Leftarrow \mathcal{O}(\mathcal{L}_1, \mathcal{F}) \Leftarrow \mathcal{O}(\mathcal{L}_2, \mathcal{F})$, that is, imposing additional structure on the loss function reduces the optimal set.

Jin, Corradi & Swanson (2017) adopt an alternative approach, based on Superiority:

$$\mathcal{S}(\mathcal{L},\mathcal{F}) : \inf_{i \in \mathcal{I}} \inf_{L \in \mathcal{L}} \mathbb{E}_{\mathcal{F}} \left[L(E_i) - L(E_M) \right] \geq 0. \tag{3}$$

The evaluated model is superior if it dominates all alternatives. In this case, the model is the only element of the optimal set; it is the 'greatest element' of the partially ordered set. Clearly, Superiority is a sufficient but not necessary condition for OPA: $\mathcal{S}(\mathcal{L},\mathcal{F}) \Rightarrow \mathcal{O}(\mathcal{L}, \mathcal{F})$.

If multiple forecasts are optimal, then the superior set is empty. The Superiority criterion therefore generally becomes non-informative as the number of forecast models increases. The OPA concept, by contrast, remains informative as the number of forecast models increases, because it always defines both (i) the (empty or singleton) superior set and (ii) the (non-empty) optimal set.

2.4 Numerical example

A random variable has a Bernoulli distribution with latent probability $\mathbb{P}[X = 1] = 0.5$. Four independent trials give rise to $2^4 = 16$ equally likely scenarios $(X_1, X_2, X_3, X_4) \in \{0, 1\}^4$. After observing the outcomes of the first three trials (X_1, X_2, X_3), three forecasts are formed for the outcome of the fourth trial (X_4): $Y_1 := \frac{1}{3}(X_1 + X_2 + X_3)$, $Y_2 := \frac{1}{8} + \frac{1}{2}Y_1$, and $Y_3 := \frac{3}{8} + \frac{1}{2}Y_1$.

It is straightforward to calculate the forecast errors $E_i = X_4 - Y_i$, $i = 1, 2, 3$, in the 16

scenarios. Forecast Y_1 is unbiased but less precise than the negatively biased forecast Y_2 and the positively biased forecast Y_3. The forecast is stochastically dominated by neither Y_2 nor Y_3, for $\mathcal{L}_1$. For example, $\mathbb{E}_{\mathcal{F}}\left[L_1^*(E_1)\right] = \frac{1}{4} < \frac{5}{16} = \mathbb{E}_F\left[L_1^*(E_2)\right]$ for $L_1^*(E) = (E)_+$; similarly, $\mathbb{E}_{\mathcal{F}}\left[L_1^{**}(E_1)\right] = \frac{1}{4} < \frac{5}{16} = \mathbb{E}_F\left[L_1^{**}(E_3)\right]$ for $L_1^{**}(E) = (-E)_+$.

Nevertheless, Y_1 does not minimize the expected value for any $L \in \mathcal{L}_1$ and, hence, it is non-optimal. For example, $\mathbb{E}_{\mathcal{F}}\left[L_1^*(E_1)\right] > \frac{3}{16} = \mathbb{E}_{\mathcal{F}}\left[L_1^*(E_3)\right]$ and $\mathbb{E}_{\mathcal{F}}\left[L_1^{**}(E_1)\right] > \frac{3}{16} = \mathbb{E}_{\mathcal{F}}\left[L_1^{**}(E_2)\right]$. To prove the universal claim, it suffices to demonstrate that there exists no feasible solution to the system of inequalities which is developed in Section 3.3.

The other two forecasts, Y_2 and Y_3, are known to be optimal, as they minimize expected loss for L_1^* and L_1^{**}, respectively. Hence, the OPA criterion reduces the choice set from three forecasts $\{Y_1, Y_2, Y_3\}$ to two forecasts $\{Y_2, Y_3\}$. By contrast, the SDS criterion does not reduce the set of forecasts, because all three forecasts are not dominated and the superior set is empty.

The example can also illustrate the power of the symmetry assumption. If the permissible loss functions are limited to $\mathcal{L}_2$, then the asymmetric loss functions L_1^* and L_1^{**} are no longer permissible and Y_1 is dominated by both Y_2 and Y_3. Therefore, also the *admissible* set is reduced to $\{Y_2, Y_3\}$, in this case.

3 Empirical Tests

3.1 Hypothesis structure

The focus is on testing a null hypothesis of optimality $(\mathcal{H}_0(\mathcal{L}, \mathcal{F}) : \mathcal{O}(\mathcal{L}, \mathcal{F}))$ versus an alternative hypothesis of non-optimality $\left(\mathcal{H}_1(\mathcal{L}, \mathcal{F}) : \mathcal{O}^C(\mathcal{L}, \mathcal{F})\right)$ for $\mathcal{L} \in \{\mathcal{L}_0, \mathcal{L}_1, \mathcal{L}_2\}$, where the superscript C denotes the logical complement. A test procedure is designed to control the test size, or frequency of false exclusions (Type I errors), by estimating critical values for a given significance level. The frequency of false inclusions (Type II errors), and thus the

test power, is not directly controlled and depends on the dimensions and structure of the data, for a given significance level.

It follows from Definition (2) that the null can be formulated as $(M-1)$ moment inequalities with respect to the infinite-dimensional parameter L:

$$\mathcal{H}_0(\mathcal{L},\mathcal{F}) : \left(\mathbb{E}_{\mathcal{F}}\left[L(E_i) - L(E_M)\right] \geq 0,\ i = 1, \cdots, M-1\right),\ L \in \mathcal{L}. \tag{4}$$

Weak inequalities are used here, because the loss function is required to exhibit non-zero variation on$\mathcal{X}_M$. The test procedure imposes this requirement by standardizing the decrements and increments of the loss function in the interior of the sample range (see Section 3.3); this standardization in turn requires the exclusion of certain irrelevant forecast models at the data pre-processing stage (see Section 3.2).

If a single given loss function L is considered, that is, $\mathcal{L} = \{L\}$, then $\mathcal{H}_0(\mathcal{L},\mathcal{F})$ reduces to the SPA hypothesis used in White (2000) and Hansen (2005). If two prospects are considered ($M = 2$), then $\mathcal{H}_0(\mathcal{L},\mathcal{F})$ reduces to a hypothesis of pairwise non-dominance as in Kaur, Prakasa Rao & Singh (1994), Davidson & Duclos (2013) and Davidson (2009), which corresponds to the alternative hypothesis of Jin, Corradi & Swanson (2017).

The null hypothesis partially identifies the loss function. The identified set is given by $\mathcal{L}^*(\mathcal{F}) := \{L \in \mathcal{L} : \mathbb{E}_{\mathcal{F}}\left[L(E_i) - L(E_M)\right] \geq 0,\ i = 1, \cdots, M-1\}$. Instead of constructing confidence sets for $\mathcal{L}^*(\mathcal{F})$, the analysis focuses on testing $\mathcal{H}_0(\mathcal{L},\mathcal{F}) : \mathcal{O}(\mathcal{L},\mathcal{F})$, which is equivalent to $\mathcal{H}_0(\mathcal{L},\mathcal{F}) : \mathcal{L}^*(\mathcal{F}) \neq \emptyset$.

The reverse hypothesis structure ($\mathcal{H}_0(\mathcal{L},\mathcal{F}) : \mathcal{O}^{\mathcal{C}}(\mathcal{L},\mathcal{F})$ vs. $\mathcal{H}_1(\mathcal{L},\mathcal{F}) : \mathcal{O}(\mathcal{L},\mathcal{F})$) seems of less interest, because non-rejection of non-optimality does not allow for exclusion of the evaluated model. In tests based on the reverse structure, false exclusions are Type II errors and beyond the control of the analyst.

3.2 Time series data

In practice, the CDF $\mathcal{F}$ is latent and has to be estimated using empirical data. Since the two empirical applications in this study use time series data, it is assumed here that the analyst has access to a discrete set of time series realizations X_t, and point forecasts $\hat{\boldsymbol{y}}_t := (\hat{y}_{1,t} \cdots \hat{y}_{M,t})$, for $t = 1, \ldots, T$.

The analysis allows for the existence of latent point forecasts $\boldsymbol{y}_t := (y_{1,t} \cdots y_{M,t})$ which are measurable functions $m_i(\boldsymbol{Z}_{i,t}, \boldsymbol{\theta}_{0_i})$ of a random vector of predictive variables, $\boldsymbol{Z}_{i,t} \in \mathbb{R}^{d_i}$, and a latent parameter vector $\boldsymbol{\theta}_{0_i} \in \text{Int}\Theta_i$ from the parameter space $\Theta_i \subseteq \mathbb{R}^{d_i}$. The forecasts at time t are constructed as $\hat{y}_{i,t} := m_i(\boldsymbol{Z}_{i,t}, \boldsymbol{\theta}_{t_i})$ for realizations $\boldsymbol{Z}_{i,t}$ and parameter estimators $\boldsymbol{\theta}_{t_i}$.

Given X_t, the unobservable error is $\boldsymbol{u}_t := X_t \mathbf{1}'_M - \boldsymbol{y}_t$ and the observed error is $\boldsymbol{\varepsilon}_t := X_t \mathbf{1}'_M - \hat{\boldsymbol{y}}_t$, where $\boldsymbol{y}_t := [m_1(\boldsymbol{Z}_{1,t}, \boldsymbol{\theta}_{1_0}) \cdots m_M(\boldsymbol{Z}_{M,t}, \boldsymbol{\theta}_{M_0})]'$ and $\hat{\boldsymbol{y}}_t := [m_1(\boldsymbol{Z}_{1,t}, \boldsymbol{\theta}_{1_t}) \cdots m_M(\boldsymbol{Z}_{M,t}, \boldsymbol{\theta}_{M_t})]'$. If m_i is independent of $\boldsymbol{\theta}_{0_i}$ and/or $\boldsymbol{\theta}_{0_i}$ is known for all i, then the original error $\boldsymbol{u}_t$ becomes observable, several of the below assumptions become obsolete, and the derivations of the limit theory become simpler.

Given the observable time series $\boldsymbol{\varepsilon}_t$, $t = 1, \ldots, T$, the latent CDF $\mathcal{F}$ is approximated by the empirical cumulative distribution function (ECDF), defined by

$$F_T(E) := T^{-1} \sum_{t=1}^{T} \mathbb{I}(\varepsilon_t \leq E). \tag{5}$$

Other CDF estimators such as those based on multivariate kernel estimation, copulas and polynomial approximations can be employed by constructing the ECDF of a large random sample generated by the relevant CDF estimator.

To simplify the exposition of the numerical procedure, it is assumed that certain forecast models are eliminated at the data pre-processing stage. For $\mathcal{L}_0$ and $\mathcal{L}_1$, models with $\min_t \varepsilon_{i,t} < \min_t \varepsilon_{M,t}$ or $\max_t \varepsilon_{i,t} > \max_t \varepsilon_{M,t}$ are excluded; for $\mathcal{L}_2$, models with $\max_t |\varepsilon_{i,t}| >$

$\max_t |\varepsilon_{M,t}|$ are excluded. These prospects do not affect the empirical OPA classification and the value of the test statistic. Unless it can be determined with sufficiently high confidence that they are non-optimal under the latent error distribution, these prospects cannot be ignored when estimating the critical values.

3.3 Empirical moment conditions

Let $\mathcal{F}_T$ be the set of all multinomial distributions with atoms at the T data points. This set includes the ECDF, that is, $F_T \in \mathcal{F}_T$. For a given $F \in \mathcal{F}_T$ with probability mass function (PMF) $f(E)$, $\mathcal{H}_0(\mathcal{L}, F)$ can be represented by a finite and exact system of linear inequalities. This system can be obtained by replacing the infinite-dimensional parameter $L \in \mathcal{L}$ by a permissible piecewise-linear loss function, along the lines of Post (2003, Thm 2). Specifically, for every permissible $L_1 \in \mathcal{L}$, there exists piecewise-linear upper envelope $L_2 \in \mathcal{L}$, such that (i) $L_1(E_M) = L_2(E_M)$ and (ii) $L_1(E_i) \leq L_2(E_i)$, $\forall i = 1, \ldots, M-1$.

It follows that the subset of piecewise-linear loss functions includes a solution to the empirical moment conditions if and only if the set $\mathcal{L}$ includes a solution. The reduction of the set of permissible loss functions therefore does not affect the truth of the empirical moment condition, the value of the test statistic or the estimated critical values which are introduced below.

For $\mathcal{L}_0$ and $\mathcal{L}_1$, let $\{z_t\}_{t=1}^{T+1}$ represent the ranked values of $\{\varepsilon_{M,t}\}_{t=1}^{T} \cup \{0\}$, so that $z_1 \leq \cdots \leq z_{T+1}$. Let $T_0 := \sup\{t : z_t < 0\}$, so that $z_{T_0+1} = 0$. For $\mathcal{L}_2$, let $\{z_t\}_{t=1}^{T+1}$ represent the ranked values of $\{|\varepsilon_{M,t}|\}_{t=1}^{T} \cup \{0\}$.

For a given General Loss function $L \in \mathcal{L}_0$, let $\beta_s := L(z_s) - L(z_{s+1}) \geq 0$, $s = 1, \cdots, T_0$, be decrements in the negative domain and $\beta_s := L(z_{s+1}) - L(z_s)$, $s = T_0+1, \cdots, T$, be increments in the positive domain. A general stepwise loss function is obtained by summation by parts:

$$L_{0,\boldsymbol{\beta}}(E) := \begin{cases} +\infty & E < z_1 \\ \sum_{s=1}^{T_0} \beta_s \mathbb{I}\left(E \leq z_{s+1}\right) + \sum_{s=T_0+1}^{T} \beta_s \mathbb{I}\left(E \geq z_s\right) & z_1 \leq E \leq z_{T+1}. \\ +\infty & E > z_{T+1} \end{cases} \tag{6}$$

Similarly, for a given Convex Loss function $L \in \mathcal{L}_1$, let $\sigma_s := \left(L(z_{s+1}) - L(z_s)\right) / \left(z_{s+1} - z_s\right)$, $s = 1, \cdots, T$, be slopes of chords between two consecutive points, and $\beta_s := \sigma_{s+1} - \sigma_s$, $s = 1, \cdots, T_0 - 1$; $\beta_{T_0} := -\sigma_{T_0}$; $\beta_{T_0+1} := \sigma_{T_0+1}$; $\beta_s := \sigma_s - \sigma_{s-1}$, $s = T_0 + 1, \cdots, T$, increments of the slopes (recall that the slope at $E = 0$ is zero). A convex piecewise-linear loss function is given by:

$$L_{1,\boldsymbol{\beta}}(E) := \begin{cases} +\infty & E < z_1 \\ \sum_{s=1}^{T_0} \beta_s \left(z_{s+1} - E\right)_+ + \sum_{s=T_0+1}^{T} \beta_s \left(E - z_s\right)_+ & z_1 \leq E \leq z_{T+1}. \\ +\infty & E > z_{T+1} \end{cases} \tag{7}$$

If a Symmetric Convex Loss function $L \in \mathcal{L}_2$ is used, then (7) reduces to

$$L_{2,\boldsymbol{\beta}}(E) := \begin{cases} \sum_{s=1}^{T} \beta_s \left(|E| - |z_s|\right)_+ & |E| \leq |z_{T+1}|. \\ +\infty & |E| > |z_{T+1}| \end{cases} \tag{8}$$

The search over the piecewise-linear functions can be performed using numerical optimization. For every forecast $i \in \mathcal{I}$ and the relevant loss function class $\mathcal{L}_j$, $j = 0, 1, 2$, define the $T \times T$ coefficient matrix $\mathbf{M}_{j,i}$ with the following elements for $s, t = 1, \cdots, T$:

$$\left(\mathbf{M}_{0,i}\right)_{t,s} := \begin{cases} \mathbb{I}\left(\varepsilon_{i,t} \leq z_{s+1}\right) - \mathbb{I}\left(\varepsilon_{M,t} \leq z_{s+1}\right) & s = 1, \cdots, T_0 \\ \mathbb{I}\left(\varepsilon_{i,t} \geq z_s\right) - \mathbb{I}\left(\varepsilon_{M,t} \geq z_s\right) & s = T_0{+}1, \cdots, T \end{cases}. \tag{9}$$

$$(\mathbf{M}_{1,i})_{t,s} := \begin{cases} (z_{s+1} - \varepsilon_{i,t})_+ - (z_{s+1} - \varepsilon_{M,t})_+ & s = 1, \cdots, T_0 \\ (\varepsilon_{i,t} - z_s)_+ - (\varepsilon_{M,t} - z_s)_+ & s = T_0{+}1, \cdots, T \end{cases}. \tag{10}$$

$$(\mathbf{M}_{2,i})_{t,s} := (|\varepsilon_{i,t}| - |z_s|)_+ - (|\varepsilon_{M,t}| - |z_s|)_+ . \tag{11}$$

The matrix $\mathbf{M}_{j,i}$ is constructed such that $\mathbf{M}_{j,i}\boldsymbol{\beta} = (L_{j,\boldsymbol{\beta}}(\varepsilon_{i,t}) - L_{j,\boldsymbol{\beta}}(\varepsilon_{M,t}))_{t=1,\cdots,T}$. The intervals where the loss function goes to infinity are ignored without harm due to the exclusion of forecast models with extreme errors (see Section 3.2). Using $\boldsymbol{p} := (f(\boldsymbol{\varepsilon}_t))_{t=1,\cdots,T}$ for the values of the PMF associated with F, it follows that $\boldsymbol{p}'\mathbf{M}_{j,i}\boldsymbol{\beta} = \mathbb{E}_F\left[L_{j,\boldsymbol{\beta}}(E_i) - L_{j,\boldsymbol{\beta}}(E_M)\right]$.

For numerical purposes, the loss function will be normalized by scalar multiplication such that $\sum_{s=1}^{T} \beta_s = 1$, without loss of generality. Combined with the non-negativity constraints, the normalization implies $\boldsymbol{\beta} \in \Delta^T$, where Δ^T is a T-simplex.

Using these arguments, $\mathcal{H}_0(\mathcal{L}_j, F)$, $j = 0, 1, 2$, is equivalent to the following linear system:

$$\begin{gathered} \boldsymbol{p}'\mathbf{M}_{j,i}\boldsymbol{\beta} \geq 0,\ i = 1, \cdots, M-1; \\ \boldsymbol{\beta} \in \Delta^T. \end{gathered} \tag{12}$$

Using duality theory for Linear Programming (LP), it is possible to obtain a similar system for testing non-optimality ($\mathcal{O}^{\mathcal{C}}(\mathcal{L}_j, F)$). Specifically, applying Farkas' lemma to (12), it is found that non-optimality occurs if and only if the evaluated forecast error distribution is dominated by some convex mixture of the other forecast error distributions, extending known results for utility functions by Bawa, Bodurtha, Rao & Suri (1985, Eq. (10)-(12), p. 423) to loss functions. Given the compelling arguments for treating $\mathcal{O}(\mathcal{L}_j, \mathcal{F})$ rather than $\mathcal{O}^{\mathcal{C}}(\mathcal{L}_j, \mathcal{F})$ as the null hypothesis, this route is not further explored here.

3.4 ELR test statistic

This study relies on a blockwise ELR test statistic, which is a transformation of a constrained non-parametric maximum log likelihood ratio.

The original time series is subdivided into $T^* := (T-B+1)$ maximally overlapping blocks of B consecutive observations, $\mathcal{B}_s := \{\boldsymbol{\varepsilon}_s, \cdots, \boldsymbol{\varepsilon}_{s+B-1}\}$, $s = 1, \cdots, T^*$. The optimal block size depends on the context and involves a trade-off between the strength of the dynamic effects and the number of independent blocks, or $\lfloor T/B \rfloor$.

Let $\mathcal{G}_T$ be the set of multinomial distributions with atoms at the T^* data blocks, represented by their PMF; let $g_T \in \mathcal{G}_T$ be the empirical probability mass functions (EPMF) of the blocks:

$$g_T(\mathcal{B}) := (T^*)^{-1} \sum_{s=1}^{T^*} \mathbb{I}\left[\mathcal{B}_s = \mathcal{B}\right]. \tag{13}$$

If $B = 1$, then $F_T(E) = \sum_{\mathcal{B}_s \leq E} g_T(\mathcal{B}_s)$, which amounts to assuming serial independence.

The block-level PMF $g \in \mathcal{G}_T$ implies an observation-level PMF f_g and CDF F_g. Specifically, observation t is included in all blocks with indices from $t^- := \max(1, t-B+1)$ to $t^+ := \min(t, T^*)$, $t = 1, \cdots, T$. Therefore, $f_g(\boldsymbol{\varepsilon}_t) \propto B^{-1} \sum_{s=t^-}^{t^+} g(\mathcal{B}_s)$, $t = 1, \cdots, T$ and $F_g(E) := \sum_{\boldsymbol{\varepsilon}_t \leq E} f_g(\boldsymbol{\varepsilon}_t)$.

Let $\mathcal{R} : (\mathcal{G}_T)^2 \to (-\infty, 0]$ be the log likelihood ratio between two multinomial block-level PMFs, so that the log empirical likelihood ratio between $g \in \mathcal{G}_T$ and g_T is given by:

$$\mathcal{R}(g, g_T) := \ln\left(\frac{\prod_{t=1}^{T^*} g(\mathcal{B}_t))}{\prod_{t=1}^{T^*} g_T(\mathcal{B}_t))}\right) = \sum_{t=1}^{T^*} \ln\left(g(\mathcal{B}_t)\right) + T^* \ln(T^*). \tag{14}$$

The constrained maximum log likelihood ratio and 'implied' probability mass function (IPMF) amount to:

$$
\begin{array}{c}
R_T(\mathcal{L}) := \sup_{g \in \mathcal{G}_T} \left\{ \mathcal{R}\left(g, g_T\right) : \mathcal{H}_0(\mathcal{L}, F_g) \right\} \\
= \sup_{L \in \mathcal{L}} \sup_{g \in \mathcal{G}_T} \left\{ \mathcal{R}\left(g, g_T\right) : \left(\mathbb{E}_{F_g}\left[L(E_i) - L(E_M) \right] \geq 0,\ i = 1, \cdots, M-1 \right),\ L \in \mathcal{L} \right\}
\end{array} ; \tag{15}
$$

$$
g_T^*(\mathcal{L}) := \arg\max_{g \in \mathcal{G}_T} \left\{ \mathcal{R}\left(g, g_T\right) : \mathcal{H}_0(\mathcal{L}, F_g) \right\}. \tag{16}
$$

The IPMF $g_T^*(\mathcal{L})$ is a constrained, non-parametric maximum likelihood estimator of the latent block-level PMF. The statistical procedure is based on the scaled ELR test statistic

$$
\mathrm{ELR}_T(\mathcal{L}) := -2 \frac{T}{T^* B} \mathcal{R}_T(\mathcal{L}). \tag{17}
$$

The scaling $\frac{T}{T^* B}$ is standard in the Blockwise EL literature (see for example (Kitamura (1997)). It approximates the number of occurrences of each individual observations inside the EL function, and it reduces to one when $B = 1$. The ELR statistic has important statistical optimality properties in the standard, point-identified case (Kitamura (2001)). Using large deviations theory, Canay (2010) concludes that inference based on the ELR statistic preserves important optimality properties, for partially identified models with moment inequality restrictions.

3.5 Statistical inference

In the Section 4, the null limit distribution of ELR_T is shown to have the form of a supremum of a potentially degenerate process with chi-bar-square marginals, under general assumptions. This insight is unfortunately of limited practical use, because the mixing weights of the process depend on the latent CDF $\mathcal{F}$, the latent exact temporal dependence

between the observations and the latent set of binding moment conditions.

Conservative statistical inference however can be based on distributions which majorize the latent limiting distribution. In a similar way, Post (2017) used the central chi-square with $(M-1)$ degrees of freedom as a general upper bound, for optimality tests based on utility functions and IID time series. This bound is reasonable when the number of models (M) is small. However, tighter bounds can be established using statistical moment selection methods in the spirit of Andrews & Jia Barwick (2012), for a larger number of models.

To implement moment selection, the present study uses the following set of forecast models which are approximately equivalent to the evaluated model for a given loss function $L \in \mathcal{L}$:

$$\mathrm{CS}(L, \mathcal{F}, c_T) := \{i = 1, \ldots, M-1 : |\mathbb{E}_{\mathcal{F}}\left[L(E_i) - L(E_M)\right]| \leq c_T\}.$$

Here, $c_T > 0$ is a sample-dependent tolerance parameter which converges to zero at an appropriate rate. This set is a superset of the 'contact set' $\mathrm{CS}(L, \mathcal{F}, 0)$ which consists of forecast models which are exactly equivalent to the evaluated model for the relevant loss function. The number of moment conditions which are approximately binding amounts to:

$$N(L, \mathcal{F}, c_T) := \#\mathrm{CS}(L, \mathcal{F}, c_T). \tag{18}$$

Critical values are estimated using a central chi-square with number of degrees of freedom equal to

$$N(L_T^*, F_{g_T^*(\mathcal{L})}, c_T) = \#\left\{i = 1, \cdots, M-1 : \left|\boldsymbol{p}^{*'}\mathbf{M}_{j,i}\boldsymbol{\beta}_T^*\right| \leq c_T\right\}, \tag{19}$$

where L_T^* is a maximizer of the empirical optimization problem in (15). Consequently, the null hypothesis is rejected at a given significance level α if and only if:

$$\mathrm{ELR}_T(\mathcal{L}) \geq q\left(1-\alpha, \chi^2_{N(L_T^*, F_{g_T^*(\mathcal{L})}, c_T)}\right), \tag{20}$$

where $q\left(1-\alpha, \chi^2_{N(L^*, F_{g^*_T(\mathcal{L})}, c_T)}\right)$ denotes the $1-\alpha$ quantile of the central chi-square distribution with degrees of freedom given in (19).

This approach is motivated by the following insights: (i) if the null is correct, then the limit distribution of $\mathrm{ELR}_T(\mathcal{L})$ is majorized by the limiting chi-bar-square of $\mathrm{ELR}_T(L_T)$; (ii) the limit distribution of $\mathrm{ELR}_T(L_T)$ in turn is majorized by the central chi-square with $N(L, \mathcal{F}, 0)$ degrees of freedom; (iii) the degrees of freedom can be estimated in an asymptotically conservative way using $N(L^*, F_{g^*_T}, c_T)$, $c_T > 0$; (iv) if the null is violated, then $\mathrm{ELR}_T(\mathcal{L})$ diverges to $+\infty$, while the quantile $q\left(1-\alpha, \chi^2_{N(L^*_T, F_{g^*_T(\mathcal{L})}, c_T)}\right)$ is almost surely bounded from above by $q\left(1-\alpha, \chi^2_M\right)$ for every T.

Section 4 demonstrates that the approach is statistically consistent and asymptotically conservative under general assumptions: the probability of a false rejection is smaller than or equal to the assumed significance level, in large samples. These results are not straightforward since the present EL framework involves moment inequalities that depend on an infinite-dimensional and partially identified parameter, and blocking arguments to account for temporal dependence.

Asymptotic conservatism can compromise the local power properties of the test. EL bootstrap methods combined with contact set estimation could be used to approximate critical values which lead to asymptotic exactness whenever the null limiting distribution is not degenerate, thereby improving asymptotic local power properties.

Existing work on the EL boostrap includes Brown & Newey (2002), Canay (2010, Section 4.1.3), Andrews & Soares (2010) and Allen, Gregory & Shimotsu (2011). However, a more general theoretical framework seems needed here , because none of the existing studies allows for the simultaneous occurrence of partial identification, parameter infinite-dimensionality, temporal dependence and blocking schemes. It is expected that a sufficiently general framework results from extending the Block Bootstrap Functional CLT of Radulovic (1996) to allow for resampling from stochastic EL-multinomial probabilities, VC-hull classes of functions (see for example Par. 2.6 in van der Vaart and Wellner, 1996) as parameters and

uniform inference.

Subsampling is an alternative approach. The results from Andrews & Guggenberger (2009) and Romano & Shaikh (2010) can be applied to show that subsampling is a uniformly valid approach to approximate the distribution of the ELR test statistic, under general sampling schemes. The main practical complication in using subsampling, in contrast to the aforementioned Block Bootstrap approach, lies in choosing the proper subsample length and compromising power in small samples due to the subsamples using only a subset of the original observations.

4 Limit Theory

The present section derives asymptotic properties for the testing procedure which was described in Section 3.5. It is assumed that the loss functions in class $\mathcal{L}_0(\mathcal{F})$ are equi-Lipschitz and that $\mathcal{F}$ is continuous, to facilitate the derivations. The Appendix includes a discussion of sufficient conditions for this assumption, situations in which it can be avoided, and reasons why it is not needed for $\mathcal{L}_1(\mathcal{F})$ and $\mathcal{L}_2(\mathcal{F})$. Section 4.1. presents and motivates the maintained statistical assumption framework. Section 4.2. derives the limiting behavior of the relevant empirical processes and the test statistic; statistical consistency is derived using the limit distribution under the null hypothesis; asymptotic conservatism is derived using the limiting behavior under the alternative hypothesis.

The main analytical challenges to obtain these results are: (i) to reduce the complexity of the loss function classes involved so as to obtain (blocking) functional limit theorems; (ii) to establish that the empirical moment conditions defined in Section 3 approximate the population moment conditions for every admissible loss function; iii) to account for the infinite dimensionality of the parameters in the moment inequality conditions, when establishing the asymptotic behavior of the EL function.

The analysis requires some additional notation. Specifically, in what follows, $\|\cdot\|$ denotes the Euclidean norm, $\ell^{\infty}(A)$ the space of real-valued bounded functions on a set A equipped

with the sup norm, and $\rightsquigarrow$ convergence in distribution. $\bar{B}_{\boldsymbol{\lambda}}(\eta)$ denotes the closed Euclidean ball in $\mathbb{R}^M$ centered at $\boldsymbol{\lambda}$ with radius equal to $\eta > 0$.

4.1 Assumption framework

The limit theory assumes that a number of conditions are satisfied for the stationarity and dependence properties of the predictive variables, smoothness properties of functions of the unknown parameters, limiting representations for the estimators of the unknown parameters and their sample sizes, the asymptotic rates of the number of blocks, as well as the slacks used in the construction of the supersets $\text{CS}(L, \mathcal{F}, c_T)$. These conditions are consistent with the empirical applications in Sections 6 and 7. The conditions which involve properties related to the latent parameters and their estimators are similar to those in McCracken (2000) and Jin, Corradi and Swanson (2017).

Assumption 4.2.1. *The following conditions hold:*

i. *For $r_T > 0$, as $T \to \infty$, $r_T \to \infty$ and $\frac{r_T}{T} \to \gamma \in (0, \infty]$.*

ii. *For all $i = 1, \ldots, M$, and any $t = 1, \ldots, T$, as $T \to \infty$, $\boldsymbol{\theta}_{i_t} = \boldsymbol{\theta}_{i_0} + H_{i_{r_T}} \left(\frac{1}{r_T} \sum_{j=t-r_T}^{t} h_{i,j} + o_{a.s.} \left(\frac{1}{\sqrt{r_T}} \right) \right)$, $H_{i_{r_T}} \rightsquigarrow H_{0_i}$ which is a non-singular $d_i \times d_i$ matrix, $\mathbb{E}[h_{i,j}] = 0_{d_i \times 1}$ and $\mathbb{E}\left[\|h_{i,j}\|^{2+\delta} \right] < +\infty$ for some $\delta > 0$.*

iii. *The vector process $\mathbf{Z}_t := \left[X_t, (\boldsymbol{Z}_{i,t}, h_{i,t})_{i=1,\ldots,M} \right]_{t \in \mathbb{Z}}$ is strictly stationary and absolutely regular with mixing coefficients $(\beta_k)_{k \in \mathbb{N}}$ that satisfy $\beta_k = O(k^{-r})$ for $r > 1$. The joint distribution of $\mathbf{Z}_0$ has continuous marginals.*

iv. *For some $\eta > 0$, such that for $\boldsymbol{\theta} := (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ restricted to $\bar{B}_{\boldsymbol{\theta}_0}(\eta) \subset \mathbb{R}^{\sum_{i=1}^{M} d_i}$, and $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{1_0}, \ldots, \boldsymbol{\theta}_{M_0})$, the function $\boldsymbol{\theta} \to u(\mathbf{Z}_0, \boldsymbol{\theta}) := X_0 \mathbf{1}'_M - [m_1(Z_{1,0}, \boldsymbol{\theta}_1) \cdots m_M(Z_{M,0}, \boldsymbol{\theta}_M)]$ is almost surely Lipschitz continuous with respect to $\boldsymbol{\theta}$, with Lipschitz coefficient $l(\mathbf{Z}_0)$,*

that satisfies $\mathbb{E}\left[l\left(\mathbf{Z}_0\right)\right] < +\infty$. *Furthermore,* $\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\bar{B}_{\boldsymbol{\theta}_0}(\eta)}\left\|u\left(\mathbf{Z}_0,\boldsymbol{\theta}\right)\right\|^p\right] < +\infty$ *for some* $p \geq 3$, *and for all* t, *the random variable* $x_t - m_M\left(Z_{M,t},\boldsymbol{\theta}_{M_t}\right)$ *has a density, that is uniformly in* t *bounded away from zero.*

v. *The functions* $\boldsymbol{\theta}_M \to \mathbb{E}_{\mathcal{F}}\left[u_M\left(\mathbf{Z}_0,\boldsymbol{\theta}_M\right)\right]$, *and* $(L,\boldsymbol{\theta}) \to \mathbb{E}_{\mathcal{F}}\left[L\left(u_i\left(\mathbf{Z}_0,\boldsymbol{\theta}\right)\right)\right]$ *are continuously differentiable with respect to* $\boldsymbol{\theta}_M$ *on* $Proj_M\bar{B}_{\boldsymbol{\theta}_0}(\eta)$, *and* $\boldsymbol{\theta}$ *on* $\bar{B}_{\boldsymbol{\theta}_0}(\eta)$, *for all* $L \in \boldsymbol{\Lambda} := \cup_{\mathbf{j}=\mathbf{0},\mathbf{1},\mathbf{2}}\boldsymbol{\mathcal{L}}_{\mathbf{j}}$ *and* $\sup_{Proj_M\bar{B}_{\boldsymbol{\theta}_0}(\eta)}\left\|D_{\boldsymbol{\theta}_M}\mathbb{E}_{\mathcal{F}}\left[u_M\left(\mathbf{Z}_0,\boldsymbol{\theta}_M\right)\right]\right\| + \sup_{\{\mathbf{1},\ldots,\mathbf{M}\}\times\boldsymbol{\Lambda}\times\bar{B}_{\boldsymbol{\theta}_0}(\eta)}\left\|D_{\boldsymbol{\theta}}\mathbb{E}_{\mathcal{F}}\left[L\left(u_i\left(\mathbf{Z}_0,\boldsymbol{\theta}\right)\right)\right]\right\| < +\infty$, *where* κ_i *denotes the* i^{th} *-coordinate of* κ *and* $Proj_i$ *denotes projection to the* i^{th} *-coordinate.*

vi. *There exists some* $\epsilon > 0$ *such that for all* $j = 0,1,2$, $\inf_{L\in\boldsymbol{\mathcal{L}}_j^*(\mathcal{F}),\ CS(L,\mathcal{F},0)\neq\emptyset}\lambda_{\min}\left(\mathcal{V}\left(L,M\right)\right) > \epsilon$, *where* $\lambda_{\min}\left(A\right)$ *denotes the minimum eigenvalue of the positive-definite matrix* A, *and* $\mathcal{V}\left(L,M\right) := \left(G_j\left(i,L\right) - G_j\left(M,L\right)\right)_{i\in CS(L,\mathcal{F},0)}$.

vii. *The block size satisfies* $B \to +\infty$ *and* $B = O\left(T^{\rho}\right)$ *for* $0 < \rho < \frac{1}{2}$.

viii. *The slacks satisfy* $c_T \to 0$ *and for any subsequence* $(T_\star)$, $\sqrt{T_\star}c_{T_\star} \to +\infty$ *almost surely.*

What follows is a discussion of the plausibility of the above conditions for typical applications.

Assumptions 4.2.1.(i)-(ii) are satisfied when the estimators of the unobserved parameters are estimated using a rolling window and the window size r_T is of the same asymptotic order as T. Assumption 4.2.1.(ii) allows for a large class of pseudo-consistent M-estimators, for example, the OLSE, GMME or Quasi MLE, which asymptotically satisfy smooth enough estimating equations, for example, under interior differentiability conditions for the criteria involved.

The first part of Assumption 4.2.1.(iii) holds for processes that satisfy a general class of stochastic recurrence equations which include GARCH models, among others; see, for example, Mikosch and Straumann (2006, Section 4). This condition implies stationarity of

the forecast errors which is not a harmless assumption. Notably, it does not allow for the recursive estimation of latent model parameters, for example, using an expanding estimation window. However, the analysis does permit a fixed, rolling or moving estimation window. The second part excludes stationary processes with univariate marginals with atoms which is a usual assumption in non-parametric inference.

The first two parts of Assumption 4.2.1.(iv) hold for predictive regressions, in which case m_i is bilinear in $(\mathbf{Z}_{i,0}, \boldsymbol{\theta}_i)$, as long as the regressors have enough moments. The final part of the assumption can be established if $\mathbf{Z}_t$ jointly with the sample upon which $\boldsymbol{\theta}_{M_t}$ depends, has a density with integral over each level set of $x_t - m_M\left(Z_{M,t}, \boldsymbol{\theta}_{M_t}\right)$ which is bounded away from zero uniformly in t due to the results in Hillier and Armstrong (1999, Section 3).

Assumption 4.2.1.(v) would follow from Assumption 4.2.1.(iii)-(iv), the uniform Lipschitz properties for the loss functions in each class (see Section 2.3 and the Appendix) and dominated convergence, if u and the loss functions involved are moreover assumed to be Lebesgue almost everywhere continuously differentiable (see, for example, Assumption A.0 in Jin, Corradi and Swanson (2017) for a similar restriction).

Assumption 4.2.1.(vii) is similar to the Definition 3.1.(iii)-(iv) of Canay (2010); it holds whenever the random vector $L\left(K\left(\mathbf{Z}_0, \boldsymbol{\theta}_0\right)\right)$ consists of linearly independent random variables, for all $L \in \mathcal{L}_j^*(\mathcal{F})$ with non-empty contact sets, and the set $\mathcal{L}_j^*(\mathcal{F})$ is compact in the topology of pointwise convergence, due to Lemma A.0 in the Appendix, and Assumption 4.2.1.(iv).

Assumption 4.2.1.(vii) employs the usual restriction on the block size divergence rate in comparison to T (see, for example, Thm.3 of Kitamura (1997)). Assumption 4.2.1.(viii) restricts the slacks so as to converge to zero at a slower rate than $\sqrt{T}$, similar to the relevant econometric literature; see Andrews and Soares (2010) and references therein.

4.2 Empirical processes, null limit theory and test properties

A theorem is established about the limiting properties of the empirical processes associated with the moment conditions, the limiting distribution for $\mathrm{ELR}_T(\mathcal{L})$, for $\mathcal{L} = \mathcal{L}_j,\ j = 0, 1, 2$, and the conservatism and consistency properties of the proposed test procedure.

Theorem 4.2.2. *Under Assumption 4.2.1.(i)-(vii) for $j = 0, 1, 2$, as $T \to \infty$:*

$$\sqrt{T}\left[\mathbb{E}_{g_T}\left[L\left(\varepsilon_{i,t}\right)\right] - \mathbb{E}_{\mathcal{F}}\left[L\left(u_{i,0}\right)\right]\right] \rightsquigarrow \mathbb{G}_j\left(i, L\right) \text{ in } \ell^\infty\left(A_j\right), \tag{21}$$

where $A_j := \{1, \ldots, M\} \times \mathcal{L}_j$, G_j, are defined by (21)-(22) with $\mathcal{L}_0$ restricted to $\mathcal{L}_0^*(\mathcal{F})$, $\mathbb{G}_j$ are zero mean Gaussian processes with covariance kernels

$$\begin{array}{c} K_{\mathbb{G}_j}\left(\left(i, L\right), \left(i^*, L^*\right)\right) := \sum_{t=0}^{\infty} \kappa_t Cov\left(L\left(u_{i,0}\right), L^*\left(u_{i^*,t}\right)\right) \\ +\varrho \sum_{t=0}^{\infty} \kappa_t Cov\left(L\left(u_{i,0}\right), D_{\boldsymbol{\theta}}\mathbb{E}_{\mathcal{F}}\left[L\left(u_i\left(\mathbf{Z}_0, \boldsymbol{\theta}_0\right)\right)\right] \boldsymbol{H}\boldsymbol{h}_t\right) \\ +\varrho \sum_{i=0}^{\infty} \kappa_t Cov\left(L^*\left(u_{i^*,0}\right), D_{\boldsymbol{\theta}}\mathbb{E}_{\mathcal{F}}\left[L^*\left(u_i\left(\mathbf{Z}_0, \boldsymbol{\theta}_0\right)\right)\right] \boldsymbol{H}\boldsymbol{h}_t\right) \\ +\varrho_\star D_{\boldsymbol{\theta}}\mathbb{E}_{\mathcal{F}}\left[L\left(u_i\left(\mathbf{Z}_0, \boldsymbol{\theta}_0\right)\right)\right] \boldsymbol{H} V_h \boldsymbol{H}' D_{\boldsymbol{\theta}}\mathbb{E}_{\mathcal{F}}\left[L^*\left(u_i\left(\mathbf{Z}_0, \boldsymbol{\theta}_0\right)\right)\right]' \end{array}, \tag{22}$$

$i, i^* \in \{1, \ldots, M\}$, $L, L^* \in \mathcal{L}_j$, *and* $\kappa_t = \begin{cases} 1, & t = 0 \\ 2, & t > 0 \end{cases}$; *in addition,* $\boldsymbol{H}$ *is the* $\sum_{i=1}^M d_i \times \sum_{i=1}^M d_i$ *block diagonal matrix* $diag_{1 \le i \le \sum_{i=1}^M d_i}\left(H_{0_i}\right)$, $\boldsymbol{h}_t := \left(h_{i,t}\right)'_{i=1,\ldots,M}$, $V_h := \sum_{t=0}^{\infty} \kappa_t \mathbb{E}\left[\boldsymbol{h}_0 \boldsymbol{h}_t'\right]$, *and*

$$\varrho = \begin{cases} 1 - \frac{\gamma}{2}, & \gamma < 1 \\ \frac{1}{2\gamma}, & \gamma \in [1, +\infty] \end{cases}, \ \varrho_\star = \begin{cases} 1 - \frac{\gamma}{3}, & \gamma < 1 \\ \frac{1}{\gamma} - \frac{1}{3\gamma^2}, & \gamma \in [1, +\infty] \end{cases}.$$

Furthermore, under $\mathcal{H}_0(\mathcal{L}_j, \mathcal{F})$ and as $T \to \infty$: i) if $\forall L \in \mathcal{L}_j^*(\mathcal{F})$, $CS(L, \mathcal{F}, 0) \neq \emptyset$,

$$ELR_T(\mathcal{L}_j) \rightsquigarrow \inf_{L \in \mathcal{L}_j^*(\mathcal{F})} \inf_{v \in \mathbb{R}_+^{N(L,\mathcal{F},0)}} \left(\mathcal{V}\left(L, M\right) - v\right)' Var^{-1}\left(\mathcal{V}\left(L, M\right)\right) \left(\mathcal{V}\left(L, M\right) - v\right), \tag{23}$$

and for i, i^* such that $\left(L, i\right), \left(L, i^*\right) \in CS(L, \mathcal{F}, 0)$ the (i, i^*) element of $Var\left(\mathcal{V}\left(L, M\right)\right)$ is given by $K_{\mathbb{G}_j}\left(\left(i, L\right), \left(i^*, L\right)\right) - K_{\mathbb{G}_j}\left(\left(i, L\right), \left(M, L\right)\right) - K_{\mathbb{G}_j}\left(\left(M, L\right), \left(i^*, L\right)\right) + K_{\mathbb{G}_j}\left(\left(M, L\right), \left(M, L\right)\right)$,

ii) if $\exists L \in \mathcal{L}_j^*(\mathcal{F})$, $CS(L, \mathcal{F}, 0) = \emptyset$,

$$ELR_T(\mathcal{L}_j) \rightsquigarrow 0. \tag{24}$$

Finally, if also Assumption 4.2.1.(viii) holds, then, for any $\alpha \in (0,1)$, *and as* $T \to \infty$,

A. Under $\mathcal{H}_0(\mathcal{L}_j, \mathcal{F})$ *if i) above holds then*

$$\limsup_{T\to\infty} \mathbb{P}\left(ELR_T(\mathcal{L}_j) \geq q\left(1-\alpha, \chi^2_{N(L^*, F_{g_T^*(\mathcal{L}_j)}, c_T)}\right)\right) \leq \alpha, \tag{25}$$

while if ii) above holds then

$$\lim_{T\to\infty} \mathbb{P}\left(ELR_T(\mathcal{L}_j) \geq q\left(1-\alpha, \chi^2_{N(L^*, F_{g_T^*(\mathcal{L}_j)}, c_T)}\right)\right) = 0. \tag{26}$$

B. Under $\mathcal{H}_1(\mathcal{L}_j, \mathcal{F})$

$$\lim_{T\to\infty} \mathbb{P}\left(ELR_T(\mathcal{L}_j) \geq q\left(1-\alpha, \chi^2_{N(L^*, F_{g_T^*(\mathcal{L}_j)}, c_T)}\right)\right) = 1. \tag{27}$$

The convergence result in (21) specifies Gaussian limits for the empirical processes associated with the moment conditions. The covariance kernels in (22) reflect the sample error variation through their first term, the estimated parameters error variation through their last term and the covariation between the two errors through the remaining terms. When $D_{\boldsymbol{\theta}}\mathbb{E}_{\mathcal{F}}\left[L\left(\kappa_i\left(\mathbf{Z}_0, \boldsymbol{\theta}\right)\right)\right] = 0_{\sum_{i=1}^M d_i}$ for all i, L, and/or $\gamma = \infty$, any (co-) variation due to the estimated parameters error disappears.

The results in (23)-(24) define the limiting distribution of the test statistic under the null. Using the Gaussian limits in (21), and arguments similar to the ones in the last part of the proof of Th. 3.1 of Canay (2010), the limit has the form of the infimum of a stochastic process over $\mathcal{L}^*(\mathcal{F})$, which has potentially degenerate chi-bar square marginals with latent weights.

Degeneracy at zero occurs if and only if the contact set is empty, or $CS(L, \mathcal{F}, 0) = \emptyset$, for some $L \in \mathcal{L}^*(\mathcal{F})$. This situation is not uncommon due to the discrete nature of the ground set of alternatives.

However, non-degeneracy is natural in several important cases. For example, if the ground set of alternatives includes nesting forecast model specifications, then the contact sets are non-empty by construction and the limit distribution is non-degenerate. Similarly, if security prices or returns are forecasted in informationally efficient markets and the ground set includes the random walk model, then the contact set will be non-empty; the random walk model in this case operates as a theoretical greatest element which enters in every contact set.

Despite its theoretical importance, the limit distribution of the test statistic cannot be used directly for statistical inference in practice, as it depends on the latent CDF $\mathcal{F}$, the latent temporal dependence between the observations, the latent set of loss functions that support the null, and the latent set of binding moment conditions.

The rejection region is constructed via the stochastic chi-square distribution introduced in Section 3.5. This distribution is not generally expected to have a weak limit in probability and may only possess sub-sequential limits. Despite this, the majorizing arguments in the construction of the critical values, as well as the restrictions on the limiting behavior of the slacks in Assumption 4.2.1.(viii) ensure that the test will be asymptotically conservative in both the degenerate and the non degenerate cases. Also, the test remains consistent under the alternative hypothesis, due to the divergence to infinity of the test statistic and the boundedness from above of the quantiles used as critical values.

4.3 Extensions for further research

A number of extensions are considered for further research.

In the framework of Assumption 4.2.1.(i)-(ii), the analysis can be extended to allow for recursive and/or fixed sampling schemes for the construction of the estimators $\boldsymbol{\theta}_{t_i}$ for some $i \in \{1, \ldots, M\}$. Using the results of West and McCracken (1998) and McCracken (2000), and extending Assumption 4.2.1.(i) and.(ii) as Assumption A.2 and A.4 of Jin, Corradi & Swan-

son (2017), (21) would continue to hold, featuring however more complicated expressions for $K_{\mathbb{G}_j}$. The analysis can also be extended to allow for cases where Θ_i for some $i \in \{1, \ldots, M\}$ is infinite dimensional, in the spirit of Linton, Song and Whang (2010).

Another possible extension is the use of side information in the form of moment conditions in addition to the optimality hypothesis. Side information about the forecasts error distribution may stem from application-specific knowledge about the forecasts models. For example, in Accounting and Finance, the sign of the forecast bias can sometimes be determined based on accounting conventions such as 'conservatism' or modeling assumptions such as 'risk neutrality'. We expect that relevant generalizations of (21) would reveal further asymptotic efficiency gains, especially if the additional moment conditions don't introduce estimation risk for additional latent parameters.

Further extensions concern the optimal choice of the slacks and the block size in finite samples, especially in the presence of parameter estimation error. A benchmark value for c_T is given in Canay (2010, Eq. 4.9). This specification could be improved in practice via tailor-made Monte Carlo simulation experiments. The block size B can be chosen by some empirical variance minimization method such as the one in El Ghouch et al. (2011).

5 Computational strategy

5.1 Auxiliary LP tests

Before computing the ELR test statistic, a number of auxiliary tests are recommended, to lower the computational burden. It is recommended to first test whether there exists a solution to the linear system (12) for the ECDF ($\boldsymbol{p} = T^{-1}\mathbf{1}_T$), using LP. If a feasible solution exists, then it follows directly that the evaluated model is optimal in the sample, $G_T^*(\mathcal{L}) = G_T$ and $ELR_T(\mathcal{L}) = 0$. If no feasible solution can be found, then the value of the

test statistic must be computed or approximated.

In the applications in Section 5 and Section 6, an LP problem is employed, to test existence of a feasible solution. The left-hand-side of the constraints in linear system (11) are augmented with positive slack variables. The objective is to minimize the sum of these slack variables. The resulting LP problem always has a feasible solution and attaining an optimal value of zero for the objective function implies that the model is fully optimal in the sample.

5.2 General problem

Let $\boldsymbol{\pi} \in \Delta^{T^*}$ be model variables which capture the block-level probabilities $(g(\mathcal{B}_t))_{t=1,\cdots,T^*}$, $G \in \mathcal{G}_T$. The associated observation-level probabilities $(f_G(\boldsymbol{\varepsilon}_t))_{t=1,\cdots,T}$ are given by $\boldsymbol{p} \propto \mathbf{P}\boldsymbol{\pi}$, where the $T \times T^*$ matrix $\mathbf{P}$ has elements $\mathbf{P}_{t,s} := B^{-1}\mathbb{I}\left(t^- \leq s \leq t^+\right)$, $t = 1, \cdots, T$; $s = 1, \cdots, T^*$. Let

$$\boldsymbol{g}_j(\boldsymbol{\beta}, \boldsymbol{\pi}) := \left(\boldsymbol{\pi}' \left(\mathbf{P}'\mathbf{M}_{j,1}\right)\boldsymbol{\beta} \cdots \boldsymbol{\pi}' \left(\mathbf{P}'\mathbf{M}_{j,M-1}\right)\boldsymbol{\beta}\right)', \; j = 0, 1, 2. \tag{28}$$

The likelihood ratio $R_T(\mathcal{L}_j)$ and the ICDF $G_T^*(\mathcal{L}_j)$ for $j = 0, 1, 2$, can be computed by solving the following optimization problem:

$$\max \mathbf{1}'_{T^*} \ln(\boldsymbol{\pi}) + T^* \ln(T^*) \tag{29}$$
$$\text{s.t. } \boldsymbol{g}_j(\boldsymbol{\beta}, \boldsymbol{\pi}) \geq \mathbf{0}_{M-1};$$
$$\boldsymbol{\beta} \in \Delta^T;$$
$$\boldsymbol{\pi} \in \Delta^{T^*}.$$

The multiplicative constraints $\boldsymbol{g}_j(\boldsymbol{\beta}, \boldsymbol{\pi}) \geq \mathbf{0}_{M-1}$ are generally not convex. However, the sub-problems for given values of $\boldsymbol{\beta} \in \Delta^T$ are standard Convex Optimization problems. Hence, the problem could be solved by enumerating a sufficiently large number of piecewise-linear candidate solutions for $\boldsymbol{\beta}$ and solving all corresponding Convex Optimization problems, along the lines of Post (2017).

5.3 Iterative strategy

Unfortunately, the number of required candidate solutions in the above approach quickly explodes as the number of forecast models increases.

A more efficient procedure recognizes that the sub-problems for given values of $\boldsymbol{\pi} \in \Delta^{T^*}$ are also standard Convex Optimization problems. The procedure alternates between (i) optimization over $\boldsymbol{\beta}$ given a solution for $\boldsymbol{\pi}$ and (ii) optimization over $\boldsymbol{\pi}$ given a solution for $\boldsymbol{\beta}$, in an iterative manner. The procedure essentially combines Generalized Methods of Moments (GMM) for estimating the loss function and EL for estimating the probabilities.

Let $\boldsymbol{\pi}_0^* = \boldsymbol{\pi^*}_1 := T^{-1}\mathbf{1}_T$ and $\boldsymbol{\pi}_t^*$, $t = 2, \cdots$, the solution to the following maximization problem:

$$\max \mathbf{1}'_{T^*} \ln(\boldsymbol{\pi}) + T^* \ln(T^*) \tag{30}$$
$$\boldsymbol{g}_j(\boldsymbol{\beta^*}_{t-1}, \boldsymbol{\pi}) \geq \mathbf{0}_{M-1};$$
$$\boldsymbol{\pi} \in \Delta^T.$$

In this problem, $\boldsymbol{\beta^*}_0 := T^{-1}\mathbf{1}_T$ and $\boldsymbol{\beta^*}_t$, $t = 1, \cdots$, is the solution to the following minimization problem:

$$\min \boldsymbol{\varepsilon}' \mathbf{W}\left(\boldsymbol{\beta}^*_{t-1}, \boldsymbol{q}_t\right) \boldsymbol{\varepsilon} \tag{31}$$
$$\boldsymbol{g}_j\left(\boldsymbol{\beta}, \boldsymbol{q}_t\right) + \boldsymbol{\varepsilon} \geq \mathbf{0}_{M-1};$$
$$\boldsymbol{\beta} \in \Delta^T;$$
$$\boldsymbol{\varepsilon} \geq \mathbf{0}_{M-1}.$$

Here, $\boldsymbol{q}_t \in \Delta^{T^*}$ is a prior solution for the probabilities based on the history $\boldsymbol{\pi^*}_s$, $s = 1, ..., t$. To avoid convergence after one iteration and, hence, allow for updating of the estimates, our application uses a specification based on a lagged moving average: $\boldsymbol{q}_t = \frac{1}{2}\left(\boldsymbol{\pi^*}_{t-1} + \boldsymbol{\pi^*}_t\right)$. The weighting matrix is set equal to the identity matrix $\mathbf{W}\left(\boldsymbol{\beta}, \boldsymbol{q}\right) = \mathbf{I}_{M-1}$, to avoid problems with estimating and manipulating the error covariance matrix.

The procedure exploits the close relationship between EL and GMM. Problem (30) is a standard EL problem with given model parameters; problem (31) amounts to an iterated GMM problem with given probabilities.

Convergence to an optimum in a finite number of iterations cannot be formally proven under general conditions. The goodness of the solutions in the empirical applications is

however supported by a high robustness to the choice of the starting values $\boldsymbol{\pi}_0^*$ and $\boldsymbol{\beta^*}_0$ and weighting matrix $\mathbf{W}\left(\boldsymbol{\beta}_{t-1}^*, \boldsymbol{q}_t\right)$, as well as a close proximity to the optimal value of the objective function which is found using the aforementioned brute-force approach (enumerating all relevant piecewise-linear loss functions) for a number of randomly selected problems.

5.4 Resampling

The computational cost of the above numerical methods would be substantial when using resampling methods, because multiple optimization problems would have to be solved for every pseudo-sample or sub-sample. This approach would require High Performance Computing or, alternatively, reducing the number of iterations, which in turn lowers the approximation precision.

In order to save computer time, it is recommended to first check whether the hypothesis of OPA can be rejected using the asymptotically conservative critical value for the desired significance level and switch to re-sampling methods only in case of non-rejection. After all, rejection at the conservative critical value already suffices to discard the evaluated forecast model given the significance level.

Since this study considers a very large application in Section 6, the focus is on the computationally friendly, conservative inference method based on a majorizing chi-squared distribution and a moment selection procedure. Encouragingly, the conservative approach proves to be sufficiently powerful to achieve very large set reductions in this application.

5.5 Hardware and software

For the empirical analysis in this study, the auxiliary LP problem and the non-linear optimization problems are modeled in GAMS. The GAMS modeling environment is deployed

within MATLAB to facilitate data handling operations. The LP problems are solved with the CPLEX solver of IBM ILOG CPLEX Optimization Studio 12.8.0.0; the non-linear problems are solved with the CONOPT4 NNLP solver (Drud (1985)).

All problems are solved on a Dell PowerEdge M610 server computer with 2 x Intel Xeon CPU E5620 processors with 2.40GHz speed and 48GB memory. The problem size is largest for the extended data set in Section 6 ($M = 3,657$ and $T = 225$). In this data set, the average solution time is approximately 100 seconds for the auxiliary LP problem and 600 seconds for the embedded Convex Optimization problems in the iterative procedure for solving the non-linear problem.

6 Forecasting Exchange Rates

A first, small-scale application replicates and extends the empirical study of exchange rate predictability of Jin, Corradi & Swanson (2017). Three forecast models are studied (3): the spot price (SP), the forward price (FP) and the three-month-lagged three-month Moving Average (MA). The original study does not include the MA forecast. This third model is added here to better illustrate the difference between the optimality and superiority criteria.

The proposed OPA test is performed for all three forecasts (SP, FP, MA) and three loss function classes ($\mathcal{L}_0$, $\mathcal{L}_1$, $\mathcal{L}_2$), for six currency pairs: Canadian dollar (CAD), French franc (FF), German mark (DM), Japanese yen (JPY), Swiss franc (CHF) and British pound (GBP), all measured against the US dollar. Daily data from Thomson Reuters Datastream are used for the sample period from January 1, 1992, to February 28, 2002. The forecasts horizon is three months. For each currency pair, the data set consists of $T \approx 2,750$ daily forecasts for each of the $M = 3$ models.

The block size of the ELR test is set at 63 days, to account for overlapping of the forecasts horizons. Since only three models are considered, the degrees of freedom for the

asymptotic chi-square test is either 1 or 2, depending on the number of alternatives which achieve approximately the same expected loss level as the evaluated model (see (19)).

Table I summarizes the test results by reporting the ELR test statistic for every combination of the three forecast models, three loss function classes and six currencies.

Using the General Loss class ($\mathcal{L}_0$), all forecasts are classified as optimal at conventional significance levels for each of the six currencies. These findings imply that none of the three forecasts is superior (for all currencies) and, moreover, every forecast is optimal for some admissible loss function and thus not redundant (for five out of six currencies) for the General Loss class.

The picture changes when the analysis is based on convex loss functions ($\mathcal{L}_1$). In this case, MA is classified as significantly non-optimal for all six currency pairs. Consequently, MA can be discarded for the $\mathcal{L}_1$ class, as the conservative nature of the test procedure ensures that the probability of a false non-optimality classification does not exceed the significance level in large samples. This result illustrates the additional discriminatory power from assuming that the loss function is convex.

A further reduction of the choice set is however not possible for most currencies. The SP and FP forecasts are both optimal and hence non-redundant for five out of six currencies (FF, DM, JPY, CHF and GBP). No forecast is superior in these cases. These findings illustrate the limitations of the superiority criterion compared with the optimality criterion.

Requiring the loss functions to be symmetric in addition to convex ($\mathcal{L}_2$) does further reduce the choice set. Specifically, for the CAD, FF, DM and CHF currencies, the SP is the unique optimal forecast for all Symmetric Convex Loss functions; for the JPY and GBP, both SP and FP are optimal. These results illustrate the incremental effect on the discriminatory power of the symmetry assumption.

Due to the small number of forecast models, the potential reduction of the choice set in this application is naturally limited to just two forecasts (leaving one non-redundant forecast); Section 5 develops a larger-scale application based on the comparison of thousands

of inflation forecasts models of Hansen (2005).

[Insert Table I about here.]

7 Forecasting US Inflation

A second, large-scale application extends the empirical study of inflation forecasts of Hansen (2005). The analysis compares thousands of distinct linear regression models which are constructed from a set of 27 predictive variables. While Hansen evaluated the regression models using a given Laplacian loss function, the present study consider entire families of loss functions (GL, CL, SCL).

Table II lists the predictive regressors and provides details about their definition and construction. Five regressors related to the Phillips Curve (Phillips (1958)) are denoted by an asterisk ($X^*_{6,t}$, $X^*_{7,t}$, $X^*_{8,t}$, $X^*_{9,t}$, $X^*_{10,t}$); these 'PC regressors' are given special attention here because of their strong predictive power in Hansen's study. The analysis considers $3,656$ distinct linear regression models with one, two or three out of the 27 regressors, and, in addition, the random walk model, a total of $M = 3,657$ models.

[Insert Table II about here.]

The analysis is performed using both Hansen's original data set and an updated data set. The two data sets are based the same set of forecasts models, but a different sample period and different vintages of the underlying data for the predictive regressors.

The original data from 1952Q1 through 1999Q4 is used to make quarterly forecasts of the end-of-quarter annual inflation change. Each regression model uses a time series of 32

quarterly observations. The first forecast is thus made at the end of 1959Q4 and uses data from 1952Q1 through 1959Q4 to predict the change in inflation between the end of 1960Q1 and the end of 1961Q1; the last forecast is made at the end of 1999Q3 for the change in inflation between the end of 1999Q4 and the end of 2000Q4. The evaluation period thus includes T= 160 quarters.

The updated data set uses the most recent vintage available from the FRED database on May 31, 2018. These data are used to generate updated series of forecasts for the original sample period and the subsequent 16-year period. The first forecast is again made at the end of 1959Q4; the final forecast is now made at the end of 2015Q4, predicting the change of inflation between the end of 2016Q1 and the end of 2017Q1.[1] The updated evaluation period thus includes T= 225 quarters.

Given the multitude of models, reduction of the choice set is highly desirable. The hypothesis of OPA is tested for each of the 3,657 forecasting model against all alternative models, for each of the three loss function families.

A blockwise application of the ELR test seems not needed in this application, as the quarterly data exhibit limited serial dependence and, furthermore, the forecast horizons are not overlapping; the block length is therefore set at $B = 1$. The number of degrees of freedom for the asymptotic chi-square test is again equal to the number of alternative models which achieve approximately the same expected loss level as the evaluated model (see (19)).

A summary overview of the frequency of non-rejection at different significant levels (α) is provided in Table III. For both data sets and several significance levels, the table shows the number of models for which OPA cannot be rejected and the fraction of such models out of the total number of models.

As shown in the table, OPA cannot be rejected at any significance level for the General Loss class for the large majority of the 3,657 models. For the original data set, 2,500 models

[1]Forecasts after 2015Q4 can not be made due to the unavailability of one of the predictive variables, namely the producer price index for finished consumer foods (PPIFCF), as explained in the relevant page of the FRED website: https://fred.stlouisfed.org/series/PPIFCF.

(68.36%) are fully GLSD optimal. The number rises to 3,641 (99.56%) when working with the updated data set. These findings illustrate the lack of discriminatory power of the OPA criterion for the General Loss loss function class.

For the Convex Loss class, only 85 forecast models (2.32%) are fully CLSD optimal, using the original data set. The set reduction from 2,500 to 85 models illustrates the power of the convexity assumption. The symmetry assumption further shrinks the choice set: only 31 models (0.85%) are SCLSD optimal. This number is even smaller than expected using the aforementioned $\sqrt{M}$ 'rule' based on existing applications of SD optimality in finance and welfare analysis, since $\sqrt{M} = \sqrt{3,657} \approx 60$ in the present application.

The results for the updated data set similarly show impressive set reductions for the Convex Loss and Symmetric Convex Loss function classes compared to the General Loss class. Only 73 models (2.00%) are fully CLSD optimal. Assuming symmetry is again very effective: only 13 models (0.36%) are SCLSD optimal for this data set.

Naturally, the optimal set expands as the significance level is lowered. However, the set reductions remain substantial. Importantly, the incremental effect of the symmetry assumption in terms of the number of exclusions is strongest for low levels of significance. At the 1% level of significance, optimality cannot be rejected for 1,410 models (38.56%) for the Convex Loss class and only 895 models (24.47%) for the Symmetric Convex Loss class.

[Insert Table III about here.]

To further diagnose the results, Logit regression analysis is performed to explain the OPA test results with variables which capture features of the evaluated forecasting models. The dependent variable D_{CFO} is a dummy which takes a value of one if the ELR test statistic equals zero. The explanatory variables are D_{PC}, or a dummy which takes a value of one when at least one PC regressor is included in the forecast model, N_{PC}, which denotes the number of included PC regressors, and N_{All}, or the total number of regressors in the predictive model.

The results of the Logit regression analysis are reported in Table IV. These results confirm Hansen's conclusion that PC variables are important predictors, for both data sets and all families of loss functions. More specifically, the statistically significant coefficient for D_{PC} demonstrates that the inclusion of PC regressors systematically increases the likelihood of forecast optimality. However, the number of PC regressors appear to matter neither for the Convex Loss class nor the Symmetric Convex Loss class, witness the insignificant role of the regressor N_{PC}. By contrast, the total number of regressors, N_{All}, does appear relevant: increasing the number of regressors systematically decreases the likelihood of forecast optimality. Overall, these results suggest that both PC regressors and model parsimony are important in forecasting inflation.

[Insert Table IV about here.]

8 Conclusion

To compare multiple forecasts in the face of ambiguity regarding the relevant loss function, the OPA hypothesis extends the SPA hypothesis by White (2000) and Hansen (2005) from a single given loss function to an entire class of loss functions.

The work by Fishburn (1974), Bawa, Bodurtha, Rao & Suri (1985) and Post (2017) was extended by (i) identifying forecast comparison as a new application area for SD Optimality; (ii) using three distinct classes of loss functions instead of utility functions; (iii) a blockwise implementation of EL; (iv) less conservative critical values using a moment selection procedure; (v) an explicit and general statistical limit theory for the test statistic and estimated critical values; (vi) a computationally more efficient computational strategy which alternates between two distinct standard Convex Optimization problems.

The earlier application of SD criteria to forecast comparison by Jin, Corradi & Swanson (2017) was extended by (i) adopting the powerful concept of optimality instead of superiority; (ii) considering the class of Symmetric Convex Loss functions in addition to General Loss and Convex Loss functions to improve discriminatory power; (iii) a hypothesis structure which allows for controlling the probability of false model rejections; (iv) a formulation in terms of moment inequality conditions which allows for efficient moment-based inference methods; (v) developing an empirical application for a very broad cross-section of forecast models.

The proposed framework was applied to the small-scale empirical study of exchange rate predictability by Jin, Corradi & Swanson (2017) and the larger study of inflation forecast models of Hansen (2005). A very large majority of thousands of inflation forecast models can be discarded for all standard loss functions. Confirming the conclusion by Hansen (2005), the optimal set consists mostly of forecast models with a Phillips Curve structure.

References

[1] Allen, J., A.W. Gregory and K. Shimotsu, 2011, Empirical likelihood block bootstrapping, *Journal of Econometrics* 161, 110-121.

[2] Anderson, G. and Th. Post, 2018, Increasing Discriminatory Power in Well-being Analysis using Convex Stochastic Dominance, forthcoming in *Social Choice and Welfare*.

[3] Andrews D. W., 1994, Empirical process methods in econometrics, *Handbook of econometrics*, 4, 2247-2294.

[4] Andrews, D.W.K. and P. Guggenberger, 2009, Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities, *Econometric Theory* 25(3), 669-709.

[5] Andrews, D.W.K. and P. Jia Barwick, 2012, Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure, *Econometrica* 80, 2805-2826.

[6] Andrews, D.W.K and G. Soares, 2010, Inference for parameters defined by moment inequalities using generalized moment selection, *Econometrica* 78(1), 119-157.

[7] Bawa, V.S., J.N. Bodurtha Jr., M.R. Rao and H.L. Suri, 1985, On Determination of Stochastic Dominance Optimal Sets, *Journal of Finance* 40, 417–431.

[8] Brown, B. and W. Newey, 2002, Generalized Method of Moments, Efficient Bootstrapping, and Improved Inference, *Journal of Business & Economic Statistics* 20, 507–517.

[9] Canay, I.A., 2010, EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity, *Journal of Econometrics*, 156 (2), 408-425.

[10] Davidson, J., 1994, *Stochastic limit theory: An introduction for econometricians*, OUP Oxford.

[11] Davidson, R., 2009, Testing for Restricted Stochastic Dominance: Some Further Results, *Review of Economic Analysis* 1, 34–59.

[12] Davidson, R. and J.-Y. Duclos, 2013, Testing for Restricted Stochastic Dominance, *Econometric Reviews* 32, 84-125.

[13] DeVore, R. A., and Lorentz, G. G., 1993, *Constructive approximation* (Vol. 303). Springer Science & Business Media.

[14] Drud, A., 1985, CONOPT: A GRG code for large sparse dynamic nonlinear optimization problems, *Mathematical Programming* 31, 153-191.

[15] El Ghouch, A., Van Keilegom, I., & McKeague, I. W., 2011, Empirical likelihood confidence intervals for dependent duration data. *Econometric Theory*, 27(1), 178-198.

[16] Fishburn, P.C., 1974, Convex stochastic dominance with continuous distribution functions, *Journal of Economic Theory* 7, 143-158.

[17] Hadar, J. and W.R. Russell, 1969, Rules for Ordering Uncertain Prospects, *American Economic Review* 59, 2–34.

[18] Hanoch, G., and H. Levy, 1969, The Efficiency Analysis of Choices Involving Risk, *Review of Economic Studies* 36, 335-346.

[19] Hansen, PR, 2005, A test for superior predictive ability, *Journal of Business and Economics Statistics* 23, 365–380.

[20] Hillier, G., & Armstrong, M.,1999, The density of the maximum likelihood estimator, *Econometrica*, 67(6), 1459-1470.

[21] Jin, S., V. Corradi and N.R. Swanson, 2017, Robust Forecast Comparison, *Econometric Theory* 33, 1306-1351.

[22] Kaur, A., B.L.S. Prakasa Rao and H. Singh, 1994, Testing for second order stochastic dominance of two distributions, *Econometric Theory* 10, 849-866.

[23] Kitamura, Y., 1997, Empirical likelihood methods with weakly dependent processes, *Annals of Statistics* 25, 2084-2102.

[24] Kitamura, Y., 2001, Asymptotic Optimality of Empirical Likelihood for Testing Moment Restrictions, *Econometrica*, 69(6), 1661–1672.

[25] Knight, K., 1999, Epi-convergence in distribution and stochastic equi-semicontinuity, *mimeo*.

[26] Linton, O.B., E. Maasoumi and Y.-J. Whang, 2005, Consistent Testing for Stochastic Dominance under General Sampling Schemes, *Review of Economic Studies* 72, 735-765.

[27] Linton, O., Song, K., & Whang, Y. J., 2010, An improved bootstrap test of stochastic dominance, *Journal of Econometrics*, 154(2), 186-202.

[28] McCracken, M.W., 2000, Robust out-of-sample inference, *Journal of Econometrics* 99, 195–223.

[29] Mikosch, T., and Straumann, D., 2006, Stable limits of martingale transforms with application to the estimation of GARCH parameters, *The Annals of Statistics*, 34(1), 493-522.

[30] Molchanov, I., 2006, *Theory of random sets*. Springer Science & Business Media.

[31] Niculescu, C., and Persson, L. E., 2006, *Convex functions and their applications* (pp. xvi+-255). New York: Springer.

[32] Owen, A., 1988, Empirical Likelihood Ratio Confidence Intervals for a Single Functional, *Biometrika* 75, 237–249.

[33] Owen, A., 1990, Empirical Likelihood for Confidence Regions, *Annals of Statistics* 18(1), 90–120.

[34] Owen, A., 1991, Empirical Likelihood for Linear Models, *Annals of Statistics* 19(4), 1725–1747.

[35] Phillips, A.W., 1958, The Relationship between Unemployment and the Rate of Change of Money Wages in the United Kingdom 1861-1957, *Economica* 25, 283–299.

[36] Pollard, D., 2010, *A User's Guide to Measure Theoretic Probability*, Cambridge University Press.

[37] Post, Th., 2003, Empirical Tests for Stochastic Dominance Efficiency, *Journal of Finance* 58, 1905-1932.

[38] Post, Th., 2017, Empirical Tests for Stochastic Dominance Optimality, *Review of Finance* 21, 793-810.

[39] Post, Th. and V. Potì, 2017, Portfolio Analysis using Stochastic Dominance, Relative Entropy and Empirical Likelihood, *Management Science* 63, 153-165.

[40] Post, T., Karabatı, S., & Arvanitis, S., 2018, Portfolio optimization based on stochastic dominance and empirical likelihood, *Journal of Econometrics*, 206(1), 167-186.

[41] Radulović, D., 1996, The bootstrap for empirical processes based on stationary observations, *Stochastic processes and their applications*, 65(2), 259-279.

[42] Radulović, D., & Wegkamp, M., 2017, An elementary proof of the weak convergence of empirical processes, *mimeo*.

[43] Rio, E., 1993, Covariance inequalities for strongly mixing processes*, Annales de l'IHP Probabilités et statistiques*, Vol. 29, No. 4, pp. 587-597.

[44] Rio, E., 2017, *Asymptotic theory of weakly dependent random processes* (Vol. 80), Berlin: Springer.

[45] Rothschild, M and JE Stiglitz, 1970, Increasing Risk: I. A Definition, *Journal of Economic Theory* 2, 225-243.

[46] Romano, J.P. and A.M. Shaikh, 2010, Inference for the Identified Set in Partially Identified Econometric Models, *Econometrica* 78, 169-211.

[47] West, K. D., and McCracken, M. W., 1998, Regression-based tests of predictive ability, *International Economic Review* 39, 817-840.

[48] White, H., 2000, A reality check for data snooping, *Econometrica* 68, 1097–1126.

[49] van der Vaart, A.W., 2000, *Asymptotic Statistics*, Cambridge University Press.

van der Vaart, A. W., and J.A. Wellner, 1996, *Weak Convergence and Empirical Processes*, Springer.

Appendix

A.1 Further representations of the stochastic orders

The empirical moment conditions in (15) depend on the random empirical atoms $\{z_t\}_{t=1}^{T+1}$. To account for the asymptotic behavior of the set of atoms in the derivation of the limit theory in Section 4, equivalent representations for the stochastic orders are established. The representations are based on loss functions which can be formulated as functionals on $\mathcal{P}(\mathcal{X}_M)$, the set of probability distributions that are supported inside $\mathcal{X}_M$.

Lemma A.0. *Given the definitions of $\mathcal{L}_j$, $j = 0, 1, 2$, we have that:*

[50] $E_M \succeq_{\mathcal{L}_0,\mathcal{F}} E_i$ *iff*

$$\mathbb{E}_{\mathcal{F}}\left[\int_{a_M}^{0} \mathbb{I}(E_i < z) - \mathbb{I}(E_M < z)\, dF(z)\right] + \mathbb{E}_{\mathcal{F}}\left[\int_{0}^{b_M} \mathbb{I}(E_i \geq z) - \mathbb{I}(E_M \geq z)\, dF(z)\right] \geq 0 \ , \tag{32}$$

for every $F \in \mathcal{P}(\mathcal{X}_M)$.

2. $E_M \succeq_{\mathcal{L}_1,\mathcal{F}} E_i$ *iff*

$$\mathbb{E}_{\mathcal{F}}\left[\int_{a_M}^{0}(z-E_i)_+-(z-E_M)_+\,dF(z)\right]+\mathbb{E}_{\mathcal{F}}\left[\int_{0}^{b_M}(E_i-z)_+-(E_M-z)_+\,dF(z)\right]\geq 0\ , \tag{33}$$

for every $F\in\mathcal{P}(\mathcal{X}_M)$.

3. $E_M \succeq_{\mathcal{L}_2,\mathcal{F}} E_i$ *iff*

$$\mathbb{E}_{\mathcal{F}}\left[\int_{\min\{a_M,-b_M\}}^{0}(|E_i|-|z|)_+-(|E_M|-|z|)_+\,dF(z)\right]\geq 0, \tag{34}$$

for every $F\in\mathcal{P}(\mathcal{X}_M)$.

Proof of Lemma A.0. For GLSD, it suffices to show that each $L\in\mathcal{L}_0$ can be characterized as $L(x)=\begin{cases}F_L(0)-F_L(x), & x<0\\ F_L(x)-F_L(0), & x\geq 0\end{cases}$, by a unique $F_L\in\mathcal{P}(\mathcal{X}_M)$, and vice versa. For the reverse implication, L can be chosen so that $L(a_M)+L(b_M)\leq 1$, as a result of the scale invariance of the inequalities $\mathbb{E}_{\mathcal{F}}[L(E_i)]\leq\mathbb{E}_{\mathcal{F}}[L(E_M)]$. Let $F_L(x):=\begin{cases}1-L(b_M)-L(x), & x<0\\ 1-L(b_M)+L(x), & x\geq 0\end{cases}$, which is a well-defined CDF supported on $\mathcal{X}_M$, due to the above restriction and the right-continuity of L (by Assumption 2.1.1). The direct implication follows from this representation. The cases of CLSD ($\mathcal{L}_1$) and SCLSD($\mathcal{L}_2$) follow from the integral representations of continuous convex functions in compact intervals, see, for example, Pollard (2010, Thm. C.3.4) along with the arguments that lead to the proof of Th.1 of Russel and Seo (1989), or Niculescu and Persson (2006, Th. 1.6.3).■

Lemma A.0 generalizes Propositions 2.2 and 2.3 of Jin, Corradi & Swanson (2017) in the sense that it characterizes the GLSD and CLSD orders without requiring almost everywhere differentiability for the loss functions involved and extends the characterization to SCLSD. Result (32) implies that if $E_M\succeq_{\mathcal{L}_0,\mathcal{F}}E_i$, then $[\mathcal{F}_M(x)-\mathcal{F}_i(x)]\operatorname{sgn}(x)\geq 0$ for every $x\in\mathcal{X}_M$, by choosing F as the degenerate distribution at x, using $\mathcal{F}_M$ and $\mathcal{F}_i$ for the CDFs of the marginal distributions of E_M and E_i, respectively. The reverse holds if $\mathcal{F}$ is assumed to be continuous, or, alternatively, Assumption A.0 of Jin, Cor-

radi & Swanson (2017) holds, due to the integration by parts property for the Lebesgue-Stieljes integral. Using the same reasoning, integration by parts, and this stricter assumption framework, it can be found that $E_i \succeq_{\mathcal{L}_1,\mathcal{F}} E_M$ iff $\int_{a_M}^{x} (\mathcal{F}_i(z) - \mathcal{F}_M(z))\, dz \mathbb{I}(x < 0) + \int_{x}^{b_M} (\mathcal{F}_M(z) - \mathcal{F}_i(z))\, dz \mathbb{I}(x \geq 0) \geq 0$. Analogously we also come up with the novel representation, $E_i \succeq_{\mathcal{L}_2,\mathcal{F}} E_M$ iff $\int_{\min\{a_M,-b_M\}}^{-|x|} (\mathcal{F}_i(z) - \mathcal{F}_M(z))\, dz + \int_{|x|}^{-\min\{a_M,-b_M\}} (\mathcal{F}_M(z) - \mathcal{F}_i(z))\, dz \geq 0$, for all $x \in [\min\{a_M, -b_M\}, 0]$.

A.2 Equi-Lipschitz properties

Part of the analysis in this study assumes that loss functions are equi-Lipschitz. The Lipschitz continuity property implies the uniform approximation of every element in the class by a piecewise linear Lipschitz function, which in turn facilitates the approximation by the loss functions implied by the empirical moment conditions in Section 3.3 the following. Furthermore, the uniformity in the Lipschitz continuity property implies that the approximation errors are independent of the loss functions involved (see Paragraph A.3). In the presence of the parameter $\boldsymbol{\theta}$, the uniformity in the Lipschitz continuity property facilitates the reduction of the metric complexity of the loss function space, so that results involving the use of entropy integrals are accessible (see for example Condition (8.33) in Th. 8.3 of Rio, 2017). Such a restriction is avoidable if results like the Decoupling Lemma of Radulović and Wegkamp (2017) can be extended to functions of uniformly bounded variation defined on multivariate Euclidean spaces. When $u(\mathbf{Z}_0, \boldsymbol{\theta})$ is independent of $\boldsymbol{\theta}$, or $\boldsymbol{\theta}_0$ is known, then this lemma is applicable and thereby this restriction is not actually needed (see also Example 2.10.7 of van der Vaart and Wellner, 1996).

For CLSD ($\mathcal{L}_1$) and SCLSD ($\mathcal{L}_2$), the uniform Lipschitz property follows from the representation result in Pollard (2010, Thm C.3.4)-see also the terms inside the expectations $\mathbb{E}_{\mathcal{F}}$ in (33)-(34) of Lemma A.0. By contrast, the equi-Lipschitz property does not generally hold for GLSD ($\mathcal{L}_0$) and an alternative justification of the property is required.

Due to DeVore and Lorentz (1993, Thm 9.3 and relation (9.10)) and the continuity of $\mathcal{F}$, the equi-Lipschitz property holds, if for example, $\mathcal{L}_0(\mathcal{F})$ is restricted to be comprised of functions which belong to a uniformly bounded subset of the Sobolev space $W_\infty^1(\mathcal{X}_M)$ (see DeVore and Lorentz (1993) Ch.2, Par. 5 for the definition and properties of the Sobolev spaces). Finally, our derivations under the weaker assumption, that $\mathcal{L}_0(\mathcal{F})$ is comprised of functions that are Lipschitz continuous, while for some $\delta > 0$, the set $\mathcal{L}_0^{*\delta}(\mathcal{F}) := \left\{L \in \mathcal{L}_0(\mathcal{F}) : \exists L^* \in \mathcal{L}_0^*(\mathcal{F}),\ \sup_{x\in\mathcal{X}_M} |L(x) - L^*(x)| \leq \delta\right\}$ is equi-Lipschitz.

A.3 Approximation by the empirical moment inequalities

Suppose that $\{z_t\}_{t=1}^{T+1}$ is as in Section 3.3. The following lemma along with the order representations of Lemma A.0, imply among others that any loss function in the relevant class, can be asymptotically uniformly in probability approximated by a loss function constructed by some appropriate (stochastic) distribution supported on $\{z_t\}_{t=1}^{T+1}$ with approximation rate $O_p\left(\frac{1}{T}\right)$. In what follows, and given the results of Lemma A.0, for $G \in \mathcal{P}(\mathcal{X}_M)$, the loss function L associated with G is defined by, $\left(\int_{a_M}^0 \mathbb{I}(x < z)\, dF(z) + \int_0^{b_M} \mathbb{I}(x \geq z)\, dF(z)\right)|_{F=G}$, or $\left(\int_{a_M}^0 (z-x)_+\, dF(z) + \int_0^{b_M} (x-z)_+\, dF(z)\right)|_{F=G}$, or $\left(\int_{\min\{a_M,-b_M\}}^0 (|x|-|z|)_+\, dF(z)\right)|_{F=G}$, and $\mathbb{E}_{\mathcal{F}}[L_T(E_i)]$ is defined by $\left(\mathbb{E}_{\mathcal{F}}\left[\int_{a_M}^0 \mathbb{I}(E_i < z)\, dF(z)\right] + \mathbb{E}_{\mathcal{F}}\left[\int_0^{b_M} \mathbb{I}(E_i \geq z)\, dF(z)\right]\right)|_{F=G}$, or $\left(\mathbb{E}_{\mathcal{F}}\left[\int_{a_M}^0 (z-E_i)_+\, dF(z)\right] + \mathbb{E}_{\mathcal{F}}\left[\int_0^{b_M} (E_i-z)_+\, dF(z)\right]\right)|_{F=G}$, or $\left(\mathbb{E}_{\mathcal{F}}\left[\int_{\min\{a_M,-b_M\}}^0 (|E_i|-|z|)_+\, dF(z)\right]\right)|_{F=G}$, for the cases $j = 0, 1, 2$ respectively.

Lemma A.1. *Suppose that $\mathcal{L}$ is equi-Lipschitz. Then the following hold:*

1. *For any $L \in \mathcal{L}$ and $\varepsilon > 0$, there exists a piecewise linear $L_\varepsilon \in \mathcal{L}$ such that $\sup_{x\in\mathcal{X}_M} |L(x) - L_\varepsilon(x)| \leq \varepsilon$.*

2. *Under Assumption 4.2.1.(i)-(v), for any $G \in \mathcal{P}(\mathcal{X}_M)$, as $T \to \infty$, w.h.p. there exists a stochastic distribution G_T supported on $\{z_t\}_{t=1}^{T+1}$, such that $\sup_{x\in\mathcal{X}_M} |L(x) - L_T(x)| = O_p\left(\frac{1}{T}\right)$, where L_T denotes the piecewise linear loss associated with G_T. Subsequently,*

$\left|\mathbb{E}_{\mathcal{F}}\left[L^k\left(E_i\right)\right]-\mathbb{E}_{\mathcal{F}}\left[L_T^k\left(E_i\right)\right]\right|=O_p\left(\frac{1}{T^k}\right)$, *for all* $i=1,\ldots,M$. *All the remainders are uniform in* L.

Proof of Lemma A.1. 1. Let ℓ denote the common Lipschitz coefficient of the class. For arbitrary $L\in\mathcal{L}$, $\varepsilon>0$, choose a partition of $\mathcal{X}_M$ with mesh equal to $\frac{\varepsilon}{2\ell}$ and such that 0 is the endpoint of some member of the partition. When $\mathcal{L}=\mathcal{L}_2$ choose a partition for $[\min\{a_M,-b_M\},\max\{-a_M,b_M\}]$ instead of $\mathcal{X}_M$, and make sure that its endpoints are symmetric around zero. Construct the piecewise linear L_ε by the relation $L\left(x\right)=L_\varepsilon\left(x\right)$ for all x that are endpoints of the partition. By construction we have that $L_\varepsilon\in\mathcal{L}$. By the fact that every $L\in\mathcal{L}$ must be Lebesgue almost everywhere differentiable with absolutely bounded derivative (by ℓ), it easy to see that L_ε is Lipschitz continuous, with coefficient bounded from above by ℓ and using mean value expansions at each element of the partition, that $\sup_{x\in\mathcal{X}_M}\left|L\left(x\right)-L_\varepsilon\left(x\right)\right|\leq\varepsilon$ due to the mesh choice. 2. The first part follows from 1, the representations of L from Lemma A.0, which imply the existence of a discrete G_T supported on the stochastic set $\{z_t\}_{t=1}^{T+1}$ and that, due to Lemma A.2 (37), the final part of Assumption 4.2.1.(iv) and Lemmas 21.4.(ii) and 21.7 of van der Vaart (2000) the mesh of the partition implied by $\{z_t\}_{t=1}^{T+1}$ is w.h.p. as $T\to\infty$, $O_p\left(\frac{1}{T}\right)$. The second part follows from the first the triangle inequality and the fact that $\mathcal{L}$ is uniformly bounded. The uniformity follows from that the remainders depend only on the common Lipschitz coefficient.■

A.4 Proofs of main results

Proof of Theorem 4.2.2. (21) is derived in Lemma A.2. Suppose that L_T is some loss function constructed by some G_T as in Lemma A.1 which converges uniformly in probability to some $L\in\mathcal{L}$. Then (21), Lemma A.1 and Skorokhod representations (justifiable by Knight

(1999)) imply that

$$\sqrt{T}\left[\mathbb{E}_{g_T}\left[L^\star\left(\varepsilon_{i,t}\right)\right]-\mathbb{E}_{\mathcal{F}}\left[L^\star\left(u_{i,0}\right)\right]\right]|_{L^\star=L_T}\overset{\text{fidi}}{\to}\mathbb{G}_j\left(i,L\right)\ ,$$

where $\overset{\text{fidi}}{\to}$ denotes fidi convergence. Tightness follows from (21), and Lemma A.1, and thereby $\sqrt{T}\left[\mathbb{E}_{g_T}\left[L^\star\left(\varepsilon_{i,t}\right)\right]-\mathbb{E}_{\mathcal{F}}\left[L^\star\left(u_{i,0}\right)\right]\right]|_{L^\star=L_T}\rightsquigarrow\mathbb{G}_j\left(i,L\right)$ in $\ell^\infty\left(A_j\right)$. Assume now that $L\in\mathcal{L}^*(\mathcal{F})$. We have that, $L_T\in\mathcal{L}^*(\mathcal{F})$ w.h.p., since due to Lemma A.1.(2), $\forall\varepsilon>0, i=1,\ldots,M-1$,

$$\begin{aligned}&\mathbb{P}\left[\mathbb{E}_{\mathcal{F}}\left[L_T\left(E_i\right)\right]-\mathbb{E}_{\mathcal{F}}\left[L_T\left(E_M\right)\right]\geq\varepsilon\right]\\&\leq\mathbb{P}\left[\mathbb{E}_{\mathcal{F}}\left[L\left(E_i\right)\right]-\mathbb{E}_{\mathcal{F}}\left[L\left(E_M\right)\right]\geq\tfrac{\varepsilon}{3}\right]+\mathbb{P}\left[\left|\mathbb{E}_{\mathcal{F}}\left[L\left(E_i\right)\right]-\mathbb{E}_{\mathcal{F}}\left[L_T\left(E_i\right)\right]\right|\geq\tfrac{\varepsilon}{3}\right]\\&+\mathbb{P}\left[\left|\mathbb{E}_{\mathcal{F}}\left[L\left(E_M\right)\right]-\mathbb{E}_{\mathcal{F}}\left[L_T\left(E_M\right)\right]\right|\geq\tfrac{\varepsilon}{3}\right]=o\left(1\right).\end{aligned}$$

Using the profile likelihood arguments and the auxiliary parameterization in Canay (2010) (Section 3 and the proof of Th. 3.1), we obtain

$$\mathrm{ELR}_T(L_T)=\max_{\lambda\leq0}2\tfrac{T}{T^*B}\sum_{r=1}^{T^*}\log\left(1+\lambda'\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right)-\bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_i\right)$$

$$=\max_{\lambda_b,\lambda_s\leq0}2\frac{T}{T^*B}\sum_{r=1}^{T^*}\log\left(1+\lambda_b'\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right)-\bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in\mathrm{CS}}+\lambda_s'\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right)-\bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\notin\mathrm{CS}}\right)$$

$$=\min_{\tau\geq0}\max_{\lambda(\tau)\in\mathbb{R}^{M-1}}2\frac{T}{T^*B}\sum_{r=1}^{T^*}\log\left(1+\lambda'\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right)-\bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_i\right)-2\frac{T}{B}\lambda'\left(\tau\right)\tau$$

$$=\min_{\tau_b,\tau_s\geq0}\max_{\lambda_b\in\mathbb{R}^{N(L,\mathcal{F},0)},\lambda_s\in\mathbb{R}^{M-1-N(L,\mathcal{F},0)}}2\frac{T}{T^*B}\sum_{r=1}^{T^*}\log\begin{pmatrix}1+\lambda_b'\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right)-\bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in\mathrm{CS}}\\+\lambda_s'\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right)-\bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\notin\mathrm{CS}}\end{pmatrix}$$

$$-2\frac{T}{B}\lambda_b'\left(\tau\right)\tau_b-2\frac{T}{B}\lambda_s'\left(\tau\right)\tau_s,\tag{35}$$

with $\bar{L}_T\left(\varepsilon^\star_{i,j}\right):=\frac{1}{B}\sum_{m=1}^{B}L_T\left(\varepsilon^\star_{i,r,m}\right)$, $\varepsilon^\star_{i,j,m}$ denoting the m^{th} element of the r^{th} block $(Z_r,\ldots,Z_{r+b-1})$, CS abbreviating $CS(L,\mathcal{F},0)$ and where the multidimensional inequalities

are interpreted point-wisely. Using (21) and Lemma A.1.(2), we obtain that uniformly w.r.t. $L \in \mathcal{L}^*(\mathcal{F})$, $\mathbb{E}_{g_T}\left[\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\notin \mathrm{CS}}\right] - \mathbb{E}_{\mathcal{F}}\left[\left(L\left(u_{i,0}\right) - L\left(u_{M,0}\right)\right)_{i\notin \mathrm{CS}}\right]$ converges to zero in probability. As in the proof of Th. 3.1 of Canay (2010), we have that for the optimizer w.r.t. τ_s,

$$\tau_{s_T} = \frac{1}{T^*}\sum_{r=1}^{T^*} \frac{\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\notin \mathrm{CS}}}{1 + \lambda_T\left(L_T\right)'\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_i} \geq \frac{\frac{1}{T^*}\sum_{r=1}^{T^*}\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\notin \mathrm{CS}}}{1 + \lambda_T\left(L_T\right)'\frac{1}{T^*}\sum_{r=1}^{T^*}\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_i},$$

due to Jensen's inequality, with $\lambda_T\left(L_T\right)$ denoting the optimizer of (35) w.r.t. λ. As in Canay (2010) we have that $\lambda_T\left(L_T\right)'\frac{1}{T^*}\sum_{r=1}^{T^*}\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_i \geq 0$. The previous along with that $\lambda'_{s_T}\left(\tau\right)\tau_{s_T} = 0$, imply that $\lambda_{s_T}\left(L_T\right)$, the optimizer for the asymptotically non-binding moment inequalities of (35), eventually equals zero w.h.p, uniformly in $\mathcal{L}^*(\mathcal{F})$. Hence w.h.p.

$$\begin{gathered}\mathrm{ELR}_T(L_T) = \max_{\lambda_b\leq 0} 2\frac{T}{T^*B}\sum_{r=1}^{T^*}\log\left(1 + \lambda'_b\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}}\right) \\ \min_{\tau_b\geq 0}\max_{\lambda_b\in\mathbb{R}^{N(L,\mathcal{F},0)}} 2\frac{T}{T^*B}\sum_{r=1}^{T^*}\log\left(1 + \lambda'_b\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}}\right) - 2\frac{T}{B}\lambda'_b\left(\tau\right)\tau_b\end{gathered} \tag{36}$$

For an arbitrary sequence $0 \leq \tau^*_{b_T} = O_p\left(T^{-\frac{1}{2}}\right)$ that is uniform in $\mathcal{L}^*(\mathcal{F})$, $\lambda_{b_T}\left(L_T, \tau^*_{b_T}\right)$, the optimizer of

$$2\frac{T}{T^*B}\sum_{r=1}^{T^*}\log\left(1 + \lambda'_b\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}} + \lambda'_s\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\notin \mathrm{CS}}\right) - 2\frac{T}{B}\lambda'_b\left(\tau\right)\tau^*_{b_T}$$

satisfies

$$\frac{1}{T^*}\sum_{r=1}^{T^*}\frac{\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}}}{1 + \lambda_{b_T}\left(L_T, \tau^*_{b_T}\right)'\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}}} - \tau^*_{b_T} = \mathbf{0}.$$

As in the proof of Lem. B.3 of Canay (2010), write $\lambda_{b_T}\left(L_T, \tau^*_{b_T}\right) = c_T a_T$ with $c_T \geq 0$ and

$\|a_T\| = 1$, and notice that due to the Cauchy- Schwarz inequality

$$\begin{aligned} 0 &= \left\| \tfrac{1}{T^*} \sum_{r=1}^{T^*} \frac{\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}}}{1+\lambda_{b_T}\left(L_T, \tau^*_{b_T}\right)'\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}}} - \tau^*_{b_T} \right\| \\ &\geq \tfrac{c_T}{T^*} a_T' \sum_{r=1}^{T^*} \frac{\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}} \left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)'_{i\in \mathrm{CS}}}{1+c_T a_T'\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}}} a_T \\ &- \left| \tfrac{1}{T^*} \sum_{j=1}^{N(L,\mathcal{F},0)} e_j' \sum_{r=1}^{T^*} \left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}} - \tau^*_{b_T} \right|, \end{aligned}$$

where e_j is the j^{th} unit vector. Thus we obtain that

$$\begin{aligned} 0 \geq c_T a_T' & \frac{\frac{1}{T^*}\sum_{r=1}^{T^*}\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}} \left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)'_{i\in \mathrm{CS}}}{1+c_T \max_r \left\|\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}}\right\|} a_T \\ - & \left| \tfrac{1}{T^*} \sum_{j=1}^{N(L,\mathcal{F},0)} e_j' \sum_{r=1}^{T^*} \left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}} - \tau^*_{b_T} \right| \end{aligned}.$$

Due to Lemma A.4, Assumption A.2.1 and Lemma A.1.(2) we obtain that $\frac{1}{T^*}\sum_{r=1}^{T^*}\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}} \left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)'_{i\in \mathrm{CS}}$ converges in probability to $\mathcal{V}(L,M)$ uniformly over $\mathcal{L}^*(\mathcal{F})$, and due to Assumption A.2.1 $\lambda_{\min}(\mathcal{V}(L,M))$ is bounded away from zero. Furthermore, due to the norm equivalence property of Euclidean spaces, Assumption 4.2.1 and Lemma A.3 we obtain that $\max_r \left\|\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}}\right\| = o_p\left(T^{-\frac{1}{2}}\right)$ with uniform remainder. Putting the previous together we have that

$$\left| \tfrac{1}{T^*} \sum_{j=1}^{N(L,\mathcal{F},0)} e_j' \sum_{r=1}^{T^*} \left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}} - \tau^*_{b_T} \right| \geq c_T a_T' \frac{\lambda_{\min}(\mathcal{V}(L,M)) + o_p(1)}{1+c_T o_p\left(T^{-\frac{1}{2}}\right)} a_T \ ,$$

hence

$$\begin{aligned} O_p\left(T^{-\frac{1}{2}}\right) = \frac{1}{\lambda_{\min}(\mathcal{V}(L,M)) + o_p(1)} & \left(\left| \tfrac{1}{T^*} \sum_{j=1}^{N(L,\mathcal{F},0)} e_j' \sum_{r=1}^{T^*} \left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}} \right| + \left\|\tau^*_{b_T}\right\| \right) \\ & \geq c_T = \left\|\lambda_{b_T}\left(L_T, \tau^*_{b_T}\right)\right\|, \end{aligned}$$

due to Lemma A.4, Lemma A.1.(2). Setting $\gamma_r := \lambda_{b_T}\left(L_T, \tau^*_{b_T}\right)'\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}}$, $r_{1,T} := -\frac{1}{T^*}\sum_{r=1}^{T^*}\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}} \frac{\gamma_r^2}{1+\gamma_r}$, and noting that due to the previous and

Lemma A.3, $\max_r |\gamma_r| = o_p(1)$ and analogously

$$\|r_{1,T}\| \leq \frac{1}{T^*}\sum_{r=1}^{T^*}\left\|\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}}\right\|^3 \frac{\left\|\lambda_{b_T}\left(L_T, \tau^*_{b_T}\right)\right\|^2}{|1+\gamma_r|} = O_p\left(T^{-1}\right),$$

uniformly in $\mathcal{L}^*(\mathcal{F})$, due to the previous and Lemma A.1.(2). Using again that $\frac{1}{T^*}\sum_{r=1}^{T^*}\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in \mathrm{CS}}\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)'_{i\in \mathrm{CS}}$ converges in probability to $\mathcal{V}(L,M)$ uniformly over $\mathcal{L}^*(\mathcal{F})$, a full rank pd matrix uniformly in the part of $\mathcal{L}^*(\mathcal{F})$ consisting of the loss functions with non-empty contact sets, while the latter has minimum eigenvalue that is bounded uniformly away from zero w.r.t. the aforementioned set, and the matrix inversion is Lipschitz in such cases, we also obtain the analogous uniform convergence in probability of the inverses. Arguing as in the final part of the proof of Theorem 3.1 of Canay (2010), bounding the relevant remainders by terms that converge in probability to zero uniformly in the aforementioned set, due to the previous, the final part of Lemma A.1.(2), Lemma A.4, Assumption 4.2.1 and the uniform boundedness of $\mathcal{L}$, it is obtained that

$$\mathrm{ELR}_T(L_T) = O_p\left(\tfrac{1}{\sqrt{T}}\right) + O_p\left(\tfrac{1}{T}\right) + \min_{\tau_b\leq 0}\left(K_T - \sqrt{T}\tau_b\right)'\mathrm{Var}^{-1}\left(\mathcal{V}(L,M)\right)\left(K_T - \sqrt{T}\tau_b\right),$$

where $K_T := \sqrt{T}\mathbb{E}_{g_T}\left[\left(L_T\left(\varepsilon_{i,t}\right) - L_T\left(\varepsilon_{M,t}\right)\right)_{i\in \mathrm{CS}}\right]$ and all the remainders are uniform in L. (21) and the CMT then imply

$$\mathrm{ELR}_T(L_T) \rightsquigarrow A(L) := \inf_{v\in\mathbb{R}_+^{N(L,\mathcal{F},0)}}\left(\mathcal{V}(L,M) - v\right)'\mathrm{Var}^{-1}\left(\mathcal{V}(L,M)\right)\left(\mathcal{V}(L,M) - v\right) \text{ in } \ell^\infty\left(\mathcal{L}_j^*(\mathcal{F})\right).$$

When $L \notin \mathcal{L}^*(\mathcal{F})$, then for $U_L := \{i = 1,2,\ldots,M-1 : \mathbb{E}_{\mathcal{F}}[L(E_i) - L(E_M)] < 0\}$

$$\mathrm{ELR}_T(L_T) \geq \max_{\lambda_u\leq 0} 2\frac{T}{T^*B}\sum_{r=1}^{T^*}\log\left(1 + \lambda_u'\left(\bar{L}_T\left(\varepsilon^\star_{i,r}\right) - \bar{L}_T\left(\varepsilon^\star_{M,r}\right)\right)_{i\in U_L}\right).$$

Due to Assumption 4.2.1 and Lemma A.4, the pseudo-consistency of $\boldsymbol{\theta}_{i_t}$ for all i and the

CMT, it follows that, for any λ_u that contains at least one strictly negative co-ordinate,

$$\begin{gathered}\frac{T}{T^*B}\sum_{r=1}^{T^*}\log\left(1+\lambda_u'\left(\bar{L}_T\left(\varepsilon_{i,r}^{\star}\right)-\bar{L}_T\left(\varepsilon_{M,r}^{\star}\right)\right)_{i\in U_L}\right)= \\ \frac{T}{T^*B}\sum_{r=1}^{T^*}\log\left(1+\lambda_u'\left(\mathbb{E}_{\mathcal{F}}\left[L(E_i)-L(E_M)\right]\right)_{i\in U_L}\right) \\ +\frac{To_p(1)}{T^*B}\sum_{r=1}^{T^*}\left(1+\lambda_u'\left(\mathbb{E}_{\mathcal{F}}\left[L(E_i)-L(E_M)\right]\right)_{i\in U_L}+o_p\left(1\right)\right)^{-1}+o_p\left(1\right)\end{gathered}$$

which due to Jensen's inequality is greater than or equal to

$$\frac{T}{T^*B}\sum_{r=1}^{T^*}\log\left(1+\lambda_u'\left(\mathbb{E}_{\mathcal{F}}\left[L(E_i)-L(E_M)\right]\right)_{i\in U_L}\right)+o_p\left(1\right)+\frac{To_p\left(1\right)}{T^{*2}B}\rightsquigarrow+\infty.$$

Hence, in this case, $\mathrm{ELR}_T(L_T)\rightsquigarrow+\infty$. The previous imply the weak epi-convergence of $\mathrm{ELR}_T(L_T)$ to $\begin{cases}A\left(L\right), & L\in\mathcal{L}^*(\mathcal{F})\\ +\infty, & L\notin\mathcal{L}^*(\mathcal{F})\end{cases}$. The results in (23)-(24) follows then from Molchanov (2006, Thm 3.4) via the use of Skorokhod representations (justifiable by Knight (1999)) since $\mathcal{L}^*(\mathcal{F})$ is compact w.r.t. the uniform metric.

Now, suppose that $N(L,\mathcal{F},0)\neq 0$ for every element of $\mathcal{L}^*(\mathcal{F})$. The compactness of $\mathcal{L}^*(\mathcal{F})$ and the continuity of $\left(\mathcal{V}\left(L,M\right)-v\right)'\mathrm{Var}^{-1}\left(\mathcal{V}\left(L,M\right)\right)\left(\mathcal{V}\left(L,M\right)-v\right)$ w.r.t. L due to (21) imply that, when (23) holds, there exist limiting optimizers. Let L^* be one of the optimizers, and suppose that, for a subsequence $(T_\star)$, $L_{T_\star}^*\rightsquigarrow L^*$. Then, uniformly w.r.t. the $i\notin\mathrm{CS}(L_*,\mathcal{F},0)$, it is found that, due to the definition of slacks and the Birkhoff's ULLN, $\mathbb{E}_{\mathcal{F}_{T_\star}}\left[L_{T_\star}(E_i)-L(E_M)\right]>c_T$, eventually, almost surely. Using (38), and Skorokhod representations, we also have that, since $\sqrt{T_\star}c_{T_\star}$ diverges to infinity almost surely, uniformly w.r.t. the $i\in\mathrm{CS}(L_*,\mathcal{F},0)$, $\left|\sqrt{T_\star}\mathbb{E}_{\mathcal{F}_{T_\star}}\left[L_{T_\star}(E_i)-L(E_M)\right]\right|\leq\sqrt{T_\star}c_T$, eventually, almost surely. The previous imply that $N(L_{T_\star},\mathcal{F},c_{T_\star})$ converges in probability to $N(L_*,\mathcal{F},0)$. Then, due to the majorization arguments in Section 3.5, (25). When $N(L,\mathcal{F},0)=0$ for some element of $\mathcal{L}^*(\mathcal{F})$, the previous implies that $\mathrm{ELR}_T(\mathcal{L})$ is eventually zero w.h.p., hence (26) follows. Finally, when the null hypothesis does not hold, then the implication of the previous is that $\mathrm{ELR}_T(\mathcal{L})$ diverges to $+\infty$, while the quantile $q\left(1-\alpha,\chi^2_{N(L_T^*,F_{g_T^*(\mathcal{L})},c_T)}\right)$ is almost surely

bounded from $q\left(1-\alpha,\chi^2_M\right)$ for every T, hence (27) follows.■

A.5 Auxiliary results

Lemma A.2. *Under Assumption 4.2.1.(i)-(vi) for $j=1,2$, as $T\to\infty$:*

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(x_t-m_M\left(\boldsymbol{Z}_{M,t},\boldsymbol{\theta}_{M_t}\right)\right)\rightsquigarrow N\left(0,v\right), \tag{37}$$

with $v\geq 0$ and

$$\sqrt{T}\left[\mathbb{E}_{F_T}\left[L\left(\varepsilon_{i,t}\right)\right]-\mathbb{E}_{\mathcal{F}}\left[L\left(u_{i,0}\right)\right]\right]\rightsquigarrow\mathbb{G}_j\left(L\right)\ \textit{in}\ \ell^\infty\left(A_j\right)\ , \tag{38}$$

where $A_j:=\{1,\ldots,M\}\times\mathcal{L}_j$, $\mathbb{G}_j$ are defined by (21)-(22) with $\mathcal{L}_0$ restricted to $\mathcal{L}_0^*\left(\mathcal{F}\right)$. If furthermore Assumption 4.2.1.(vii) holds, then (21) also holds.

Proof of Lemma A.2. Assumption 4.2.1.(iii)-(iv) and Rio (2017, Cor 4.1) imply that as $T\to\infty$,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[u_M\left(\mathbf{Z}_t,\boldsymbol{\theta}_M\right)-\mathbb{E}_{\mathcal{F}}\left[u_M\left(\mathbf{Z}_0,\boldsymbol{\theta}_M\right)\right]\right]\overset{\text{fidi}}{\to}\mathcal{G}^\star\left(\boldsymbol{\theta}_M\right),\ \text{in}\ \ell^\infty\left(\text{Proj}_M\bar{B}_{\boldsymbol{\theta}_0}\left(\eta\right)\right),$$

where $\overset{\text{fidi}}{\to}$ denotes fidi convergence and u_M denotes the M^{th} element of u, while $\mathcal{G}^\star\left(\boldsymbol{\theta}\right)$ is a zero-mean Gaussian process with uniformly continuous sample paths in $\ell^\infty\left(\text{Proj}_M\bar{B}_{\beta_0}\left(\eta\right)\right)$. Define the non-negative measure $Q:=\left(1+4\sum_{k=0}^{\infty}\beta_k\right)P$, where P denotes the law of $\mathbf{Z}_0$. Notice that the the Assumption 4.2.1.(iv) implies that there exists a $C_M>0$ independent of $\boldsymbol{\theta}$, such that for any $\delta>0$,

$$\mathbb{E}_Q\left[\sup_{\boldsymbol{\theta}_{M_1},\boldsymbol{\theta}_{M_2}\in\text{Proj}_M\bar{B}_{\boldsymbol{\theta}_0}(\eta),\left\|\boldsymbol{\theta}_{M_1}-\boldsymbol{\theta}_{M_2}\right\|\leq\delta}\left|u_M\left(\mathbf{Z}_0,\boldsymbol{\theta}_{M_1}\right)-u_M\left(\mathbf{Z}_0,\boldsymbol{\theta}_{M_2}\right)\right|\right]\leq C_M\delta,$$

where $\mathbb{E}_Q$ denotes expectation w.r.t. Q. This then implies (see for example the proof of Th.

5 of Andrews (1994)) that the entropy integral condition (8.33) in Th. 8.3 of Rio (2016) holds, which along with Assumption 4.2.1.(iii) implies the applicability of this theorem to the empirical process $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[u_M\left(\mathbf{Z}_t,\boldsymbol{\theta}_M\right)-\mathbb{E}_{\mathcal{F}}\left[u_M\left(\mathbf{Z}_0,\boldsymbol{\theta}_M\right)\right]\right]$ and thereby its tightness. Hence,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[u_M\left(\mathbf{Z}_t,\boldsymbol{\theta}_M\right)-\mathbb{E}_{\mathcal{F}}\left[u_M\left(\mathbf{Z}_0,\boldsymbol{\theta}_M\right)\right]\right]\rightsquigarrow\mathcal{G}^{\star}\left(\boldsymbol{\theta}_M\right),\text{ in }\ell^{\infty}\left(\mathrm{Proj}_M\bar{B}_{\boldsymbol{\theta}_0}\left(\eta\right)\right). \tag{39}$$

Assumption 4.2.1.(i)-(ii) and (39) then imply that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[u_M\left(\mathbf{Z}_t,\boldsymbol{\theta}_{M_t}\right)-\mathbb{E}_{\mathcal{F}}\left[u_M\left(\mathbf{Z}_0,\boldsymbol{\theta}_{M_t}\right)\right]\right]\rightsquigarrow\mathcal{G}^{\star}\left(\boldsymbol{\theta}_{M_0}\right). \tag{40}$$

Now, Assumption 4.2.1.(i),(ii),(v) and the Mean Value Theorem imply that as $T\to\infty$, almost surely,

$$\begin{array}{c}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[\mathbb{E}_{\mathcal{F}}\left[u_M\left(\mathbf{Z}_0,\boldsymbol{\theta}_{M_t}\right)\right]-\mathbb{E}_{\mathcal{F}}\left[u_M\left(\mathbf{Z}_0,\boldsymbol{\theta}_{M_0}\right)\right]\right]\\ =\frac{1}{\sqrt{R_TT}}\sum_{t=1}^{T}\left[D_{\boldsymbol{\theta}_M}\mathbb{E}_{\mathcal{F}}\left[u_M\left(\mathbf{Z}_0,\boldsymbol{\theta}^{\star}_{M_t}\right)\right]\right]\sqrt{R_T}\left(\boldsymbol{\theta}_{M_t}-\boldsymbol{\theta}_{M_0}\right)\end{array},$$

with $\boldsymbol{\theta}^{\star}_{M_t}$ a random point on the ray that connects $\boldsymbol{\theta}_{M_t}$ and $\boldsymbol{\theta}_{M_0}$ inside $\mathrm{Proj}_M\bar{B}_{\boldsymbol{\theta}_0}\left(\eta\right)$. Due to Assumption 4.2.1.(i)-(v), and Van Der Vaart (2000, Thm 18.14), jointly with (39),

$$\frac{1}{\sqrt{R_TT}}\sum_{t=1}^{T}\left[D_{\boldsymbol{\theta}_M}\mathbb{E}_{\mathcal{F}}\left[u_M\left(\mathbf{Z}_0,\boldsymbol{\theta}\right)\right]\right]\sqrt{R_T}\left(\boldsymbol{\theta}_{M_t}-\boldsymbol{\theta}_{M_0}\right)\rightsquigarrow\mathcal{G}_{\star}\left(\boldsymbol{\theta}_M\right),\text{ in }\ell^{\infty}\left(\mathrm{Proj}_M\bar{B}_{\boldsymbol{\theta}_0}\left(\eta\right)\right) \tag{41}$$

where now $\mathcal{G}_{\star}\left(\boldsymbol{\theta}_M\right)$ is a zero-mean Gaussian process with uniformly continuous sample paths in $\ell^{\infty}\left(\mathrm{Proj}_M\bar{B}_{\boldsymbol{\theta}_0}\left(\eta\right)\right)$. The definition of $\boldsymbol{\theta}^{\star}_{M_t}$ and (41) then imply that jointly with (40),

$$\frac{1}{\sqrt{R_TT}}\sum_{t=1}^{T}\left[D_{\boldsymbol{\theta}_M}\mathbb{E}_{\mathcal{F}}\left[u_M\left(\mathbf{Z}_0,\boldsymbol{\theta}^{\star}_{M_t}\right)\right]\right]\sqrt{R_T}\left(\boldsymbol{\theta}_{M_t}-\boldsymbol{\theta}_{M_0}\right)\rightsquigarrow\mathcal{G}_{\star}\left(\boldsymbol{\theta}_{M_0}\right). \tag{42}$$

Eq. (37) then follows via the Continuous Mapping Theorem.

The equi-Lipschitz property of $\mathcal{L}$ and Assumption 4.2.1.(iv) implies that there exists a $C > 0$ independent of $L, \boldsymbol{\theta}$, such that for any $\delta > 0$,

$$\mathbb{E}_Q \left[\sup_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \bar{B}_{\boldsymbol{\theta}_0}(\eta), \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \leq \delta} \left| L\left(u\left(\mathbf{Z}_0, \boldsymbol{\theta}_1\right)\right) - L\left(u\left(\mathbf{Z}_0, \boldsymbol{\theta}_2\right)\right) \right| \right] \leq C\delta.$$

As previously, this implies that the entropy integral condition (8.33) in Th. 8.3 of Rio (2016) holds, which along with the uniform boundedness of $\mathcal{L}$ and Assumption 4.2.1.(iii) implies the applicability of this theorem to the empirical process $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[L\left(u\left(\mathbf{Z}_t, \boldsymbol{\theta}\right)\right) - \mathbb{E}_{\mathcal{F}}\left[L\left(u\left(\mathbf{Z}_0, \boldsymbol{\theta}\right)\right)\right]\right]$. The limiting Gaussianity follows by Rio (2017, Cor 4.1). An analogous analysis to the one leading to (41)-(42) then implies (42) via Assumption 4.2.1.(i)-(v). The form of the covariance kernel of the limit process follows by taking into account West and McCracken (1998, Lemmata 4.1-2) via Rio (2017, Cor 4.1).

Working as in the proof of Lemma 2 of El Ghouch et al. (2011), it is found that, for any $L \in \mathbb{L}$,

$$\mathbb{E}_{g_T}\left[L\left(\varepsilon_{i,t}\right)\right] - \mathbb{E}_{\mathcal{F}}\left[L\left(u_{i,0}\right)\right] = \frac{T}{T-B+1}\left(\mathbb{E}_{F_T}\left[L\left(\varepsilon_{i,t}\right)\right] - \mathbb{E}_{\mathcal{F}}\left[L\left(u_{i,0}\right)\right]\right) - \frac{1}{T-B+1}\left(U_{1,B} + U_{2,B}\right),$$

with

$$U_{1,B} := B^{-1} \sum_{j=1}^{B} (B-j)\left(L\left(\varepsilon_{i,j}\right) - \mathbb{E}_{\mathcal{F}}\left[L\left(u_{i,0}\right)\right]\right),$$

and

$$U_{2,B} := B^{-1} \sum_{j=1}^{B} (B-j)\left(L\left(\varepsilon_{i,T-j+1}\right) - \mathbb{E}_{\mathcal{F}}\left[L\left(u_{i,0}\right)\right]\right).$$

$\mathcal{L}$ can be chosen as uniformly bounded, hence, from (38) and uniform integrability,

$$\sqrt{T}\left(\mathbb{E}\left[L\left(\varepsilon_{i,t}\right)\right] - \mathbb{E}_{\mathcal{F}}\left[L\left(u_{i,0}\right)\right]\right) = o\left(1\right),$$

uniformly in L, and, thereby, due to Assumption 4.2.1.(vii), $\frac{1}{\sqrt{T}}\mathbb{E}\left(U_{1,T}\right) = o\left(1\right)$ uniformly in

L. Now, due to the uniform boundedness of $\mathcal{L}$ and

$$\frac{1}{TB^2}\mathrm{Var}\left(\sum_{j=1}^{B}(B-j)\left(L\left(\varepsilon_{i,j}\right)-\mathbb{E}_{\mathcal{F}}\left[L\left(u_{i,0}\right)\right]\right)\right)$$

$$=\frac{1}{TB^2}\sum_{j,j'=1}^{B}\left((B-j)\left(B-j'\right)\mathrm{Cov}\left(L\left(\varepsilon_{i,j}\right),L\left(\varepsilon_{i,j'}\right)\right)+o\left(\frac{1}{\sqrt{T}}\right)\right)$$

$$=\frac{1}{TB^2}\sum_{j,j'=1}^{B}(B-j)\left(B-j'\right)\mathrm{Cov}\left(L\left(\varepsilon_{i,j}\right),L\left(\varepsilon_{i,j'}\right)\right)+o\left(1\right),$$

where the remainder term is $o\left(1\right)$ uniformly in L. Then, due to Assumption 4.2.1.(i)-(iii), the absolute value of the first term in the previous display is almost surely less than or equal to

$$\frac{\sup_{\boldsymbol{\theta}\in\bar{B}_{\boldsymbol{\theta}_0}(\eta)}}{TB^2}\left|\sum_{j,j'=1}^{B}(B-j)\left(B-j'\right)\mathrm{Cov}\left(L\left(K_i\left(\mathbf{Z}_j,\boldsymbol{\theta}\right)\right),L\left(K_i\left(\mathbf{Z}_{j'},\boldsymbol{\theta}\right)\right)\right)\right|$$

and due to stationarity the latter is less than or equal to

$$\mathrm{B}\sup_{\boldsymbol{\theta}\in\bar{B}_{\boldsymbol{\theta}_0}(\eta)}\mathrm{Var}\left(L\left(K_i\left(\mathbf{Z}_0,\boldsymbol{\theta}\right)\right)\right)+\sup_{\boldsymbol{\theta}\in\bar{B}_{\boldsymbol{\theta}_0}(\eta)}\mathrm{B}\sum_{j=1}^{B}\left|\mathrm{Cov}\left(L\left(K_i\left(\mathbf{Z}_0,\boldsymbol{\theta}\right)\right),L\left(K_i\left(\mathbf{Z}_j,\boldsymbol{\theta}\right)\right)\right)\right|=\mathrm{o}(T),$$

uniformly in L, due to the uniform boundedness of $\mathcal{L}$, uniform integrability, the proof of (38), and Davydoff's inequality (see Rio, 1993). An analogous result holds for $U_{2,B}$. Then (21) follows from (38) by noting that $T\sim T-B+1$.■

Lemma A.3. *Under Assumption 4.2.1.(i)-(vii) for $j=1,2$ and in addition Assumption 4.2.1.(ix) for $j=0$, as $T\to\infty$ and for any $p>0$:*

$$\max_{1\leq s\leq T-B+1}\sup_{L\in\mathcal{L},\boldsymbol{\theta}\in\bar{B}_{\boldsymbol{\theta}_0}(\eta)}\left\|\left(B^{-1}\sum_{s=1}^{B}L\left(u\left(\mathbf{Z}_{k,j}^{\star},\boldsymbol{\theta}\right)\right)\right)_i\right\|=O_p\left(T^{\frac{1}{p}}\right),\tag{43}$$

with $\mathbf{Z}_{k,j}^{\star}$ denoting the k^{th} element of the j^{th} block $(Z_j,\ldots,Z_{j+b-1})$, and $L\left(u\right)$ denoting $\left(L\left(u_i\right)\right)_{i=1,\ldots,M}$.

Proof of Lemma A.3. As in the proof of Lemma 2 of El Ghouch et al. (2011), there exists

some $C > 0$ such that

$$\max_{1\leq j\leq T-B+1}\sup_{L\in\mathcal{L},\boldsymbol{\theta}\in\bar{B}_{\boldsymbol{\theta}_0}(\eta)}\left\|B^{-1}\sum_{k=1}^{B}L\left(u\left(\mathbf{Z}_{k,j}^{\star},\boldsymbol{\theta}\right)\right)\right\|\leq C\max_{1\leq t\leq T}\sup_{L\in\mathcal{L},\boldsymbol{\theta}\in\bar{B}_{\boldsymbol{\theta}_0}(\eta)}\left\|L\left(u\left(\mathbf{Z}_t,\boldsymbol{\theta}\right)\right)\right\|,$$

and, thus, for any $\epsilon > 0$,

$$\mathbb{P}\left[\max_{1\leq t\leq T}\sup_{L\in\mathcal{L},\boldsymbol{\theta}\in\bar{B}_{\boldsymbol{\theta}_0}(\eta)}\left\|L\left(u\left(\mathbf{Z}_t,\boldsymbol{\theta}\right)\right)\right\|>\epsilon T^{\frac{1}{p}}\right]\leq\sum_{t=1}^{T}\mathbb{P}\left[\sup_{L\in\mathcal{L},\boldsymbol{\theta}\in\bar{B}_{\boldsymbol{\theta}_0}(\eta)}\left\|L\left(u\left(\mathbf{Z}_t,\boldsymbol{\theta}\right)\right)\right\|>\epsilon T^{\frac{1}{p}}\right]$$

$$\leq\frac{1}{\epsilon^{p+\delta}T^{\frac{\delta}{p}}}\mathbb{E}\left[\left(\sup_{L\in\mathcal{L},\boldsymbol{\theta}\in\bar{B}_{\boldsymbol{\theta}_0}(\eta)}\left\|L\left(u\left(\mathbf{Z}_t,\boldsymbol{\theta}\right)\right)\right\|\right)^{p+\delta}\right]=o\left(1\right),$$

due to stationarity, the inequality of Markov, and the uniform boundedness of $\mathcal{L}$, which implies that for any $p,\delta>0$, $\mathbb{E}\left[\left(\sup_{L\in\mathcal{L},\boldsymbol{\theta}\in\bar{B}_{\boldsymbol{\theta}_0}(\eta)}\left\|L\left(u\left(\mathbf{Z}_t,\boldsymbol{\theta}\right)\right)\right\|\right)^{p+\delta}\right]<+\infty$.■

Lemma A.4. *Suppose $(X_i^n)_{1\leq i\leq n,n\in\mathbb{N}^\star}$ is a "row-wise" strictly stationary and strongly mixing triangular array of random variables with mixing coefficients $(\alpha_i^n)_{1\leq i\leq n,n\in\mathbb{N}^\star}$, such that for some $\delta>1$, $\frac{1}{n}\sum_{i=1}^{n}\left(\alpha_i^n\right)^{1-\frac{1}{\delta}}\to 0$ as $n\to\infty$. Furthermore, $\mathcal{M}$ is a totally bounded (w.r.t. the uniform metric) set of bounded real functions on $\mathbb{R}\times\Theta$, that is also equi-Lipschitz w.r.t. Θ, with Lipschitz coefficients independent of the first argument, Θ is a non-empty compact subset of some Euclidean space, and $\mathbb{E}\left[m\left(X_1^n,\theta\right)\right]\to L_{m,\theta}$ as $n\to\infty$ and $L_{m,\theta}\in\mathbb{R}$ for all $m\in\mathcal{M}$ and $\theta\in\Theta$. Then for any $\varepsilon>0$,*

$$\mathbb{P}\left[\sup_{m,\theta}\left|\frac{1}{n}\sum_{i=1}^{n}m\left(X_i^n,\theta\right)-L_{m,\theta}\right|>\varepsilon\right]=o\left(1\right).$$

Proof of Lemma A.4. The total boundedness of $\mathcal{M}$ implies that $Y_{i,n}:=\sup_{m,\theta}\left|m\left(X_i^n,\theta\right)-\mathbb{E}\left[m\left(X_1^n,\theta\right)\right]\right|$ is a well defined uniformly bounded random variable (see Par. 1.7 of van der Vaart and Wellner, 2000). The equi-Lipschitz property of $\mathcal{M}$ and the independence of the Lipschitz coefficients of the first argument, implies that $\mathbb{E}\left[m\left(X_1^n,\theta\right)\right]$

is equicontinuous (in n) w.r.t. θ. This and the point-wise convergence of $\mathbb{E}\left[m\left(X_1^n,\theta\right)\right]$ imply that $\sup_{m,\theta}\left|\mathbb{E}\left[m\left(X_1^n,\theta\right)\right]-L_{m,\theta}\right|^2\to 0$ as $n\to\infty$. Th. 14.1 of Davidson (1994) implies that the array $\left(Y_i^n\right)_{1\leq i\leq n,n\in\mathbb{N}^\star}$ is "row-wise" strictly stationary and strongly mixing with the same mixing coefficients. Then for arbitrary $\varepsilon>0$,

$$\begin{aligned}\mathbb{P}\left[\sup_{m,\theta}\left|\tfrac{1}{n}\sum_{i=1}^n m\left(X_i^n,\theta\right)-L_{m,\theta}\right|>\varepsilon\right] &\leq \frac{\mathbb{E}\left[\sup_{m,\theta}\left(\sum_{i=1}^n m\left(X_i^n,\theta\right)-L_{m,\theta}\right)^2\right]}{\varepsilon^2 n^2}\\ &\leq \frac{2\mathrm{Var}\left[\sum_{i=1}^n Y_{i,n}\right]}{\varepsilon^2 n^2}+\frac{2\sup_{m,\theta}\left|\mathbb{E}\left[m\left(X_1^n,\theta\right)\right]-L_{m,\theta}\right|^2}{\varepsilon^2}\\ \leq \frac{2\sum_{i=1}^n \mathrm{Cov}[Y_{i,n},Y_{1,n}]}{\varepsilon^2 n}+o\left(1\right) &\leq \frac{C\sum_{i=1}^n\left(\alpha_i^n\right)^{1+\delta}}{\varepsilon^2 n}+o\left(1\right)=o\left(1\right),\end{aligned}$$

where the first inequality in the previous display follows from the inequality of Markov, the second from the inequality of Jensen, the third from stationarity and the definition of $L_{m,\theta}$, and the fourth from Davydoff's inequality (see Rio, 1993) and the fact that $Y_{1,n}$ is a bounded random variable, with $C=24\sup_n\mathbb{E}\left[\left|Y_{1,n}\right|^{2\delta}\right]^{\frac{1}{\delta}}$.■

Table I: Evaluating Exchange Rate Forecast Models. The table shows the ELR test statistic for every combination of the three forecast models (SP, FP, MA), three loss function classes (GL, CL, SCL) and six currencies (CAD, FF, DM, JPY, CHF, GBP). Daily data from Thomson Reuters Datastream are used for the sample period from January 1, 1992, to February 28, 2002. The forecasts horizon is three months. The block size of the ELR test is set at 63 days. The number of degrees of freedom for the asymptotic chi-square test is either 1 or 2, depending on the number of models for which optimality cannot be rejected with near certainty. Asterisks are used to indicate the level of significance: 0.10 (*), 0.05 (**) or 0.01 (***).

Class	Model	CAD	FF	DM	JPY	CHF	GBP
	SP	0.00	0.00	0.00	0.00	0.00	0.00
GL	FF	0.00	0.00	0.00	0.00	0.00	0.00
	MA	0.00	0.00	0.36	0.00	0.00	0.84
	SP	0.00	0.00	0.00	0.00	0.00	0.00
CL	FF	132.67***	0.00	1.54	0.00	0.00	0.00
	MA	8.21***	2.94*	7.96**	3.49*	8.34***	21.37***
	SP	0.00	0.00	0.00	0.00	0.00	0.00
SCL	FF	132.67***	4.30**	8.24***	0.00	7.48***	0.00
	MA	11.72***	7.79***	20.12***	13.07***	14.90***	27.04***

Table II: Predictive Regressors. Shown are details about the predictive regressors which are used to forecast annual US inflation change. Y_t denotes the dependent variable and $X_{i,t}$, $i \in 1, 2, ..., 27$, are the regressors. The meaning of the acronyms follows: GDPCTPI = gross domestic product chain-type price index; CBI = change in private inventories; GDP = gross domestic product; TB3MS = 3-month Treasury bill rate, secondary market**; PPIENG = producer price index, fuel and related products and power***; PPIFCF = producer price index, finished consumer foods***; MANEMP = employees on nonfarm payrolls, manufacturing; Q = quarter.

Variable	Description	
Y_t	Ann inflation	$Y_t = log(GDPCTPI_t) - log(GDPCTPI_{t-4})$
$X_{1,t}, X_{2,t}$	Ann inflation (lags of Y_t)	$X_{1,t} = Y_{t-5}$, $X_{2,t} = Y_{t-8}$
$X_{3,t}, X_{4,t}$	Qtly inflation	$X_{3,t} = 4 \times [\log(GDPCTPI_t) - \log(GDPCTPI_{t-1})]$, $X_{4,t} = X_{3,t-1}$
$X_{5,t}$	Qtly inflation rt last year's inflation	$X_{5,t} = \log(1 + x_{3,t}) - \log(1 + x_{3,t-1})$
$X^*_{6,t}, X^*_{7,t}$	Chg in empl in manufacturing sector	$X_{6,t} = \log(MANEMP_t) - \log(MANEMP_{t-1})$, $X_{7,t} = X_{6,t-1}$
$X^*_{8,t}$	Qtly empl rt to avg of previous yr	$X_{8,t} = \log(MANEMP_t) - \log\left(\frac{1}{4}\sum_{i=1}^{4} MANEMP_{t-i}\right)$
$X^*_{9,t}$	Qtly empl rt avg of previous 2 yrs	$X_{9,t} = \log(MANEMP_t) - \log\left(\frac{1}{8}\sum_{i=1}^{8} MANEMP_{t-i}\right)$
$X^*_{10,t}, X_{11,t}$	Qtly chg in real inventory	$X_{10,t} = \log(CBI_t) - \log(GDP_t)$, $X_{11,t} = X_{10,t-1}$
$X_{12,t}, X_{13,t}$	Qtly chg in qtly GDP	$X_{12,t} = \log(GDP_t) - \log(GDP_{t-1})$, $X_{13,t} = X_{12,t-1}$
$X_{14,t}$	Interest paid on 3-mo T-bill	$X_{14,t} = TB3MS_t$
$X_{15,t}, X_{16,t}$	Changes in 3-mo T-bill	$X_{15,t} = \Delta X_{14,t}$, $X_{16,t} = X_{15,t-1}$
$X_{17,t}, X_{18,t}$	Changes in 3-mo T-bill r.t. level of T-bill	$X_{17,t} = \Delta X_{14,t}/X_{14,t-1}$, $X_{18,t} = X_{17,t-1}$
$X_{19,t}, X_{20,t}$	Changes in prices of food and energy	$X_{19,t} = \log(PPIENG_t) - \log(PPIENG_{t-1})$, $X_{20,t} = X_{20,t-1}$
$X_{21,t}, X_{22,t}$	Chg in prices of food	$X_{21,t} = \log(PPIFCF_t) - \log(PPIFCF_{t-1})$, $X_{22,t} = X_{21,t-1}$
$X_{23,t}, X_{24,t}, X_{25,t}, X_{26,t}$	Qtly dummies: Q1, Q2, Q3, Q4	$X_{23,t} = \mathbb{I}(Q_t = 1)$, $X_{24,t} = X_{23,t-1}$, $X_{25,t} = X_{23,t-2}$, $X_{26,t} = X_{23,t-3}$
$X_{27,t}$	Constant	$X_{27,t} = 1$

* These variables are motivated by the Philips Curve and hence are referred to as "PC variables".

** Quarterly data are defined as the average of the monthly observation over the quarter.

*** Quarterly data are defined as the last monthly observation of the quarter.

Table III: Forecast Optimality Classification. The table shows the number of SD optimal models and the percentage of such models out of the total number of models, for both the original data set and the updated data set and the three loss function families: General Loss (GL), Convex loss (CL) and Symmetric Convex Loss (SCL). Models are classified as optimal if OPA is not rejected at the given significance level (α).

Panel A: 1961Q1-2000Q4								
	$\alpha = 1.00$		$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
GL	2,500	68.36%	2,500	68.36%	2,500	68.36%	2,500	68.36%
CL	85	2.32%	1,020	27.89%	1,188	32.48%	1,532	41.89%
SCL	31	0.85%	632	17.28%	763	20.86%	1,007	27.54%

Panel B: 1961Q1-2017Q1								
	$\alpha = 1.00$		$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
GL	3,641	99.56%	3,641	99.56%	3,641	99.56%	3,641	99.56%
CL	73	2.00%	994	27.18%	1,132	30.95%	1,410	38.56%
SCL	13	0.36%	531	14.52%	658	17.99%	895	24.47%

Table IV: Logit Regression. Shown are the estimates of Logit regressions for a dummy variable which takes a value of one when the forecasts model is fully SD optimal in sample (the ELR test statistic equals zero). The explanatory variables are D_{PC}, or a dummy that takes a value of one when there is at least one PC regressor, N_{PC}, which is the number of PC regressors, and N_{All}, denoting the total number of regressors. The t-statistics are computed based on heteroskedasticity-adjusted standard errors. Also shown are the McFadden pseudo R-squared and the fraction of correctly classified cases (Cqt). Results are shown for both the original data set and the updated data set and for the three loss function families: General Loss (GL), Convex loss (CL) and Symmetric Convex Loss (SCL).

Panel A: 1961Q1-2000Q4						
	Var	Coeff	t-stat	p-value	R^2	Cqt
GL	D_{PC}	0.52	2.93	0.003		
	N_{PC}	-0.32	-2.35	0.019		
	N_{All}	-1.11	-10.20	0.000		
					0.030	0.687
CL	D_{PC}	2.62	4.04	0.000		
	N_{PC}	-0.18	-0.39	0.697		
	N_{All}	-1.72	-6.92	0.000		
					0.150	0.984
SCL	D_{PC}	1.23	2.57	0.010		
	N_{PC}	-0.03	-0.09	0.931		
	N_{All}	-0.75	-2.75	0.006		
					0.040	0.978

Panel B: 1961Q1-2017Q1						
	Var	Coeff	t-stat	p-value	R^2	Cqt
GL	D_{PC}	N/A	N/A	N/A		
	N_{PC}	N/A	N/A	N/A		
	N_{All}	N/A	N/A	N/A		
					N/A	N/A
CL	D_{PC}	1.37	2.76	0.006		
	N_{PC}	0.28	0.80	0.422		
	N_{All}	-0.71	-2.44	0.015		
					0.064	0.980
SCL	D_{PC}	1.76	1.54	0.123		
	N_{PC}	0.28	0.37	0.713		
	N_{All}	-1.05	-1.52	0.129		
					0.072	0.996