Nonparametric Tests for Superior Predictive Ability

Thierry Post[*], Valerio Potì[†]& Selcuk Karabati[‡]

September 19, 2018

Abstract

A nonparametric method for comparing multiple forecast models is developed and implemented. The hypothesis of Nonparametric Forecast Optimality 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 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 using a blockwise Empirical Likelihood Ratio test. 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.

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) rules. Their hypothesis of Stochastic Dominance Superiority (SDS) 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 SDS 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 (GL) functions and Convex Loss (CL) 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 Nonparametric Forecast Optimality (NFO).

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 (SCL) functions is introduced. The additional assumption of symmetry improves the discriminatory power upon the analysis based on GL and CL functions. The SCL class includes the standard quadratic and absolute 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 NFO 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 NFO.

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) and Post (2017) in the context of welfare analysis and asset pricing.

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 stated 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. 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 a Convex Optimization (CO) problem for estimation the loss function given the probabilities and a CO 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. Consistent critical values can be estimated with additional computational effort using statistical resampling methods.

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 NFO, provides the empirical specification of moment conditions implied by the null hypothesis of NFO and illustrates the testing procedure. Section 4 provides details on a computational strategy that can be followed to carry out the test. Section 5 provide an illustrative application to exchange rates forecasting, extending the empirical section in Jin, Corradi & Swanson (2017). Section 6 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 7 concludes.

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.

Instead of a point estimate for a random variable, Y could also represent an aggregated goodness-of-fit measure of a given forecast model for a given sample or, alternatively, a divergence measure of an estimated distribution from a latent target distribution. This approach however introduces the risk of specification error for the relevant measure.

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 GL functions, $\mathcal{L}_0$, contains all loss functions which achieve a minimum at $L(0) = 0$ and do not decrease as the error moves away from zero. The subset of CL functions $\mathcal{L}_1 \subset \mathcal{L}_0$ assumes also convexity: $wL(E_1) + (1-w)L(E_2) \geq L(wE_1 + (1-w)E_2)$. The SCL functions $\mathcal{L}_2 \subset \mathcal{L}_1$ furthermore exhibit symmetry: $L(E) = L(|E|)$.

The SCL 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 $\mathbb{E}_{\mathcal{F}}\left[L(E)\right]$ is equivalent to the maximization of $\mathbb{E}_{\mathcal{F}}\left[U\left(-\left|E\right|\right)\right] = -\mathbb{E}_{\mathcal{F}}\left[L\left(\left|E\right|\right)\right]$, where U is an increasing and concave utility function. NFO 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.

2.2 Admissibility, Optimality and Superiority

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 $\mathbb{E}_{\mathcal{F}}\left[L(E_i)\right] \leq \mathbb{E}_{\mathcal{F}}\left[L(E_M)\right]$ for all $L \in \mathcal{L}$; non-dominance occurs if $\mathbb{E}_{\mathcal{F}}\left[L(E_i)\right] > \mathbb{E}_{\mathcal{F}}\left[L(E_M)\right]$ for some $L \in \mathcal{L}$.

The distinction between strict and weak inequality is inconsequential for the present analysis, and $\mathbb{E}_{\mathcal{F}}\left[L(E_i)\right] \geq \mathbb{E}_{\mathcal{F}}\left[L(E_M)\right]$ can replace $\mathbb{E}_{\mathcal{F}}\left[L(E_i)\right] > \mathbb{E}_{\mathcal{F}}\left[L(E_M)\right]$ 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 $\mathbb{E}_{\mathcal{F}}\left[L(E_M)\right] = 0$ and thus $\mathbb{E}_{\mathcal{F}}\left[L(E_i)\right] \geqslant \mathbb{E}_{\mathcal{F}}\left[L(E_M)\right]$ would become trivial.

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:

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

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.

NFO occurs if the evaluated model minimizes expected loss for some permissible loss function:

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

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

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 SDS:

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

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, SDS is a sufficient but not necessary condition for NFO: $\mathcal{S}(\mathcal{L},\mathcal{F}) \Rightarrow \mathcal{O}(\mathcal{L}, \mathcal{F})$.

If multiple forecasts are optimal, then the superior set is empty. The SDS criterion therefore generally becomes non-informative as the number of forecast models increases. The NFO 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.

2.3 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, $\mathbb{E}_{\mathcal{F}}[L_1^*(E_1)] = \frac{1}{4} < \frac{5}{16} = \mathbb{E}_F[L_1^*(E_2)]$ for $L_1^*(E) = (E)_+$; similarly, $\mathbb{E}_{\mathcal{F}}[L_1^{**}(E_1)] = \frac{1}{4} < \frac{5}{16} = \mathbb{E}_F[L_1^{**}(E_3)]$ 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, $\mathbb{E}_{\mathcal{F}}[L_1^*(E_1)] > \frac{3}{16} = \mathbb{E}_{\mathcal{F}}[L_1^*(E_3)]$ and $\mathbb{E}_{\mathcal{F}}[L_1^{**}(E_1)] > \frac{3}{16} = \mathbb{E}_{\mathcal{F}}[L_1^{**}(E_2)]$. 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 NFO 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 reduced to $\mathcal{L}_2$, the optimal set remains $\{Y_2, Y_3\}$, as both forecasts are optimal for $L_2^* = |E|$. In this case, 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. The admissible set is therefore reduced to $\{Y_2, Y_3\}$.

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) depends on the test power.

It follows from Definition (2) that the null can be formulated as $(M-1)$ moment inequalities:

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

Weak inequalities are allowed here, because the loss function is required to be increasing on $\mathcal{X}_M$ as the forecast error moves away from zero. 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 (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 $\mathcal{L}^*(\mathcal{F}):=\{L\in\mathcal{L}:\ \mathbb{E}_{\mathcal{F}}\left[L(E_i)-L(E_M)\right]>0,\ i=1,\cdots,M-1\}$. 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 the analyst has access to a discrete time series of realizations x_t, $\boldsymbol{y}_{t-1} := (y_{1,t-1} \cdots y_{M,t-1})$, $\boldsymbol{\varepsilon}_t := x_t \mathbf{1}'_M - \boldsymbol{y}_{t-1}$, $t = 1, ..., T$. The time-series is assumed to be a strictly stationary sequence which obeys the usual regularity conditions for the blockwise EL method.

Stationarity of the forecast errors 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.

Using the data set, the empirical cumulative distribution function (ECDF) is constructed: $F_T(E) := T^{-1} \sum_{t=1}^{T} \mathbb{I}\left(\varepsilon_t \leq E\right)$. 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, 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 NFO classification and the value of the test statistic. Unless it can be determined with sufficiently high confidence that they are non-optimal, however, these prospects can not be ignored when estimating the critical values.

3.3 Empirical moment conditions

Let $\mathcal{F}_T$ be the set of all multinomial CDFs 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 $f(E)$, $\mathcal{H}_0(\mathcal{L}, F)$ 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 piecewise-linear loss function, along the lines of Post (2003, Thm 2).

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 GL 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:

$$L_{0,\boldsymbol{\beta}}(E) := \begin{cases} +\infty & E < z_1 \\ \sum_{s=1}^{T_0} \beta_s \mathbb{I}\left(E \leq z_{s+1}\right) + \sum_{s=T_0+1}^{T} \beta_s \mathbb{I}\left(E \geq z_s\right) & z_1 \leq E \leq z_{T+1}. \\ +\infty & E > z_{T+1} \end{cases} \tag{5}$$

Similarly, for a given CL 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:

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

If a SCL function $L \in \mathcal{L}_2$ is used, then (6) reduces to

$$L_{2,\boldsymbol{\beta}}(E) := \begin{cases} \sum_{s=1}^{T} \beta_s \left(|E| - |z_s|\right)_+ & |E| \leq |z_{T+1}| \,. \\ +\infty & |E| > |z_{T+1}| \end{cases} \tag{7}$$

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 $s, t = 1, \cdots, T$:

$$\left(\mathbf{M}_{0,i}\right)_{t,s} := \begin{cases} \mathbb{I}\left(\varepsilon_{i,t} \leq z_{s+1}\right) - \mathbb{I}\left(\varepsilon_{M,t} \leq z_{s+1}\right) & s = 1, \cdots, T_0 \\ \mathbb{I}\left(\varepsilon_{i,t} \geq z_s\right) - \mathbb{I}\left(\varepsilon_{M,t} \geq z_s\right) & s = T_0{+}1, \cdots, T \end{cases} . \tag{8}$$

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

$$\left(\mathbf{M}_{2,i}\right)_{t,s} := \left(|\varepsilon_{i,t}| - |z_s|\right)_+ - \left(|\varepsilon_{M,t}| - |z_s|\right)_+ . \tag{10}$$

The matrix $\mathbf{M}_{j,i}$ is constructed such that $\mathbf{M}_{j,i}\boldsymbol{\beta} = \left(L_{j,\boldsymbol{\beta}}(\varepsilon_{i,t}) - L_{j,\boldsymbol{\beta}}(\varepsilon_{M,t})\right)_{t=1,\cdots,T}$. 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 $\boldsymbol{p} := \left(f(\boldsymbol{\varepsilon}_t)\right)_{t=1,\cdots,T}$ for the values of the probability mass function associated with F, it follows that $\boldsymbol{p}'\mathbf{M}_{j,i}\boldsymbol{\beta} = \mathbb{E}_F\left[L_{j,\boldsymbol{\beta}}(E_i) - L_{j,\boldsymbol{\beta}}(E_M)\right]$.

Replacement of $L \in \mathcal{L}_j$ with $L_{j,\boldsymbol{\beta}}$ is allowed, because it preserves the loss scores of the evaluated forecast model ($L(\varepsilon_{M,t}) = L_{j,\boldsymbol{\beta}}(\varepsilon_{M,t})$) and does not decrease the scores of the alternatives ($L(\varepsilon_{i,t}) \leq L_{j,\boldsymbol{\beta}}(\varepsilon_{i,t})$ for all $i \in \mathcal{I}$).

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:

$$\begin{gathered} \boldsymbol{p}'\mathbf{M}_{j,i}\boldsymbol{\beta} \geq 0,\ i = 1, \cdots, M-1; \\ \boldsymbol{\beta} \in \Delta^T. \end{gathered} \tag{11}$$

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 (11), 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 $\mathcal{G}$ be the CDF for data blocks and $\mathcal{G}_T$ the set of multinomial CDFs with atoms at the T^* data blocks. Let $G_T \in \mathcal{G}_T$ be the empirical cumulative distribution function (ECDF) of the blocks and

$$g_T(\mathcal{B}) := (T^*)^{-1} \sum_{s=1}^{T^*} \mathbb{I}\left[\mathcal{B}_s = \mathcal{B}\right] \tag{12}$$

the associated histogram estimator of the joint density. If $B = 1$, then $\mathcal{G} = \mathcal{F}$, $\mathcal{G}_T = \mathcal{F}_T$ and $G_T = F_T$, which amounts to assuming serial independence.

Let $F_G \in \mathcal{F}_T$ the observation-level CDF which is implied by block-level CDF $G \in \mathcal{G}_T$. Observation t is included in all blocks with indices from $t^- := \max(1, t - B + 1)$ to $t^+ := \min(t, T^*)$, $t = 1, \cdots, T$. The observation-level probabilities therefore amount to $f_G(\boldsymbol{\varepsilon}_t) \propto B^{-1} \sum_{s=t^-}^{t^+} g(\mathcal{B}_s)$, $t = 1, \cdots, T$.

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

$$\mathcal{R}(G, G_T) := \ln\left(\frac{\prod_{t=1}^{T^*} g(\mathcal{B}_t))}{\prod_{t=1}^{T^*} g_T(\mathcal{B}_t))}\right) = \sum_{t=1}^{T^*} \ln\left(g(\mathcal{B}_t)\right) + T^* \ln(T^*). \tag{13}$$

The constrained maximum log likelihood ratio and ‘implied’ cumulative distribution function (ICDF) amount to:

$$R_T(\mathcal{L}) := \max_{G \in \mathcal{G}_T} \left\{\mathcal{R}(G, G_T) : \mathcal{H}_0(\mathcal{L}, F_G)\right\}; \tag{14}$$

$$G_T^*(\mathcal{L}) := \arg\max_{G \in \mathcal{G}_T} \left\{\mathcal{R}(G, G_T) : \mathcal{H}_0(\mathcal{L}, F_G)\right\}. \tag{15}$$

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

$$ELR_T(\mathcal{L}) := -2\mathcal{R}_T(\mathcal{L}). \tag{16}$$

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 is optimal also for partially identified models with moment inequality restrictions. The test statistic is expected to be more efficient than the Kolmogorov-Smirov and Cramér-von Mises type test statistics which are typically used for pairwise dominance tests.

3.5 Statistical inference

The limit distribution of $ELR_T(\mathcal{L})$ is known to be chi-bar-square, under general regularity conditions. This insight unfortunately is of limited practical use, because the mixing weights depend on the latent CDF $\mathcal{F}$. Conservative statistical inference however can be based on distributions which majorize the latent chi-bar-square.

Post (2017) used the central chi-square with $(M-1)$ degrees of freedom as a general upper bound. 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.

The number of moment conditions which are approximately binding for a given loss function $L \in \mathcal{L}$ is determined as follows:

$$N(L, \mathcal{F}, c_T) := \#\left\{i = 1, \cdots, M-1 : \left|\mathbb{E}_{\mathcal{F}}\left[L(E_i) - L(E_M)\right]\right| \leq c_T\right\}. \tag{17}$$

In this expression, $c_T > 0$ is a sample-dependent tolerance parameter which converges to zero at an appropriate rate.

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

$$N(L^*, F_{G^*_T(\mathcal{L})}, c_T) = \#\{i = 1, \cdots, M-1 : |\boldsymbol{p}^{*\prime}\mathbf{M}_{j,i}\boldsymbol{\beta}^*| \leq c_T\}. \tag{18}$$

This approach is motivated by the following insights: (i) the limit distribution of $ELR_T(\mathcal{L})$ is majorized by the limiting chi-bar-square of $ELR_T(L)$ for any $L \in \mathcal{L}$; (ii) the limit distribution of $ELR_T(L)$ 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 $N(L^*, F_{G_T^*}, c_T)$, $c_T > 0$.

This approach is conservative in the sense that the test size is smaller than the significance level. The frequency of false model exclusions (Type I errors) will thus be under control in large samples. The flip side of using conservative critical values is that the test power is compromised; false model inclusions (Type II errors) will occur more frequently than in case of exact critical values.

Statistical re-sampling methods can be used to estimate consistent critical values, to further improve the test power. The EL bootstrap of Brown & Newey (2002) generates random pseudo-samples from the ICDF. The ICDF satisfies the Golden Rules of the bootstrap: it is a consistent and efficient estimator of the null distribution and it obeys the null hypothesis (by construction).

This bootstrap method was originally developed for moment equalities, point-identified parameters and serially independent data. Canay (2010, Section 4.1.3) and Andrews & Soares (2010) propose modifications to account for inequalities and partial identification. The EL block bootstrap of Allen, Gregory & Shimotsu (2011) allows for a class of stationary dynamic processes.

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 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 Computational strategy

4.1 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 (11) 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.

4.2 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

$$\boldsymbol{g}_j(\boldsymbol{\beta}, \boldsymbol{\pi}) := \left(\boldsymbol{\pi}'\left(\mathbf{P}'\mathbf{M}_{j,1}\right)\boldsymbol{\beta} \cdots \boldsymbol{\pi}'\left(\mathbf{P}'\mathbf{M}_{j,M-1}\right)\boldsymbol{\beta}\right)', \ j = 0, 1, 2. \tag{19}$$

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:

$$\begin{gathered}\max \mathbf{1}'_{T^*} \ln(\boldsymbol{\pi}) + T^* \ln(T^*) \\ \text{s.t. } \boldsymbol{g}_j(\boldsymbol{\beta}, \boldsymbol{\pi}) \geq \mathbf{0}_{M-1}; \\ \boldsymbol{\beta} \in \Delta^T; \\ \boldsymbol{\pi} \in \Delta^{T^*}.\end{gathered} \tag{20}$$

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 CO 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 CO problems, along the lines of Post (2017).

4.3 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 CO 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:

$$\begin{gathered}\max \mathbf{1}_{T^*}' \ln(\boldsymbol{\pi}) + T^* \ln(T^*) \\ \boldsymbol{g}_j(\boldsymbol{\beta^*}_{t-1}, \boldsymbol{\pi}) \geq \mathbf{0}_{M-1}; \\ \boldsymbol{\pi} \in \Delta^T. \end{gathered} \tag{21}$$

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:

$$\begin{gathered}\min \boldsymbol{\varepsilon}' \mathbf{W}\left(\boldsymbol{\beta}^*_{t-1}, \boldsymbol{q}_t\right) \boldsymbol{\varepsilon} \\ \boldsymbol{g}_j\left(\boldsymbol{\beta}, \boldsymbol{q}_t\right) + \boldsymbol{\varepsilon} \geq \mathbf{0}_{M-1}; \\ \boldsymbol{\beta} \in \Delta^T; \\ \boldsymbol{\varepsilon} \geq \mathbf{0}_{M-1}. \end{gathered} \tag{22}$$

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 (21) is a standard EL problem with given model parameters; problem (22) 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.

4.4 Resampling

The computer burden of the re-sampling methods is substantial, because multiple optimization problems 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.

To save computer time, it is therefore recommended to first check whether the hypothesis of NFO 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.

4.5 Hardware and software

For the empirical analysis in Section 5 and Section 6, 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 times is approximately 100 seconds for the auxiliary LP problem and 600 seconds for the embedded CO problems in the iterative procedure for solving the non-linear problem.

5 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 NFO 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-squared test is either 1 or 2, depending on the number of models for which optimality cannot be rejected with near certainty.

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 GL 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 GL 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 SCL 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.]

6 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 loss function (absolute loss), 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

[1] Forecasts after 2015Q4 can not be made due to the unavailability of one of the predictive variables,

period thus includes T= 225 quarters.

Given the multitude of models, reduction of the choice set is highly desirable. The hypothesis of NFO 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-squared test is again equal to the number of models for which optimality cannot be rejected with near certainty.

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 NFO cannot be rejected and the fraction of such models out of the total number of models.

As shown in the table, NFO cannot be rejected at any significance level for the GL class for the large majority of the 3,657 models. For the original data set, 2,500 models (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 NFO criterion for the GL loss function class.

For the CL 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 CL and SCL function classes compared to the GL class. Only 73 models (2.00%) are fully CLSD

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.

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 CL class and only 895 models (24.47%) for the SCL class.

[Insert Table III about here.]

To further diagnose the results, Logit regression analysis is performed to explain the NFO 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 CL class nor the SCL 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.]

7 Conclusion

To compare multiple forecasts in the face of ambiguity regarding the relevant loss function, the NFO 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) a computationally more efficient computational strategy which alternates between two distinct standard CO 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 SCL functions in addition to GL and CL 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.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.

[4] 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.

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

[6] 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.

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

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

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

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

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

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

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

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

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

[16] Imbens, G., R.H. Spady and P. Johnson, 1998, Information Theoretic Approaches to Inference in Moment Condition Models, *Econometrica* 66(2), 333–357.

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

[18] 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.

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

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

[21] 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.

[22] Newey, W. K., and R. J. Smith, 2004, Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators, *Econometrica* 72(1), 219–255.

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

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

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

[26] 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.

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

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

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

[30] Qin, J. and J. Lawless, 1994, Empirical likelihood and general estimating equations, *Annals of Statist*ics 22, 300–325.

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

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

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

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-squared 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
GL	SP	0.00	0.00	0.00	0.00	0.00	0.00
	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
CL	SP	0.00	0.00	0.00	0.00	0.00	0.00
	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***
SCL	SP	0.00	0.00	0.00	0.00	0.00	0.00
	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 (GL. CL and SCL). Models are classified as optimal if NFO 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 (GL. CL and 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