

Robust Optimization of Forecast Combinations

November 30, 2018

Abstract

A methodology is developed for constructing robust combinations of time-series forecast models which improve upon a given benchmark specification for all symmetric and convex loss functions. The optimal forecast combination asymptotically almost surely dominates the benchmark and, in addition, optimizes the chosen goal function, under standard regularity conditions. The optimum in a given sample can be found by solving a Convex Optimization problem. An application to forecasting of changes of the S&P 500 Volatility Index shows that robust optimized combinations improve significantly upon the out-of-sample forecasting accuracy of simple averaging and unrestricted optimization.

Key words: Forecast combinations, Stochastic Dominance, Asymptotic theory, Convex Optimization, Volatility index forecasting, Time-series analysis

JEL code: C44, C54, C61.

1 Introduction

A body of literature starting with Bates and Granger (1969) and Newbold and Granger (1974) aims to improve forecast accuracy by combining multiple individual forecast models. The main purpose is to provide a combined model that performs better than the best individual model by combining the complementary information contained in the individual models. Clemen (1989) provides a review and annotated bibliography of the early work on aggregation in the area of forecasting.

Aggregation methods have been proposed also for other statistical estimation problems such as parametric and non-parametric regression estimation and density estimation; see, for example, Juditsky and Nemirovski (2000), Rigollet and Tsybakov (2007), Dalayan and Salmon (2012) and Lavancier and Rochet (2016).

Forecast combinations with estimated optimal weights often trail simple averages out of sample, due to estimation error; see Smith and Wallis (2009), Claeskens, Magnus, Vasnev and Wang (2016) and Chan and Pauwels (2018), among others. The present study addresses this problem using a methodology for constructing robust time-series forecast combinations which stochastically dominate a given benchmark specification. Natural choices for the benchmark include a constant, a random walk model, the best individual forecast and the simple average of all models.

Stochastic Dominance (SD; Hadar and Russell (1969), Hanoch and Levy (1969) and Rothschild and Stiglitz (1970)) was originally developed to compare risky choice alternatives using utility functions. Jin, Corradi and Swanson (2017) extend the use of this stochastic order to forecast comparison based on loss functions. An SD relation arises if a given forecast achieves a lower expected loss than a second forecast for all permissible loss functions, allowing for a robust classification.

Whereas Jin, Corradi and Swanson (2017) test for dominance relations among a set of given forecast models in the spirit of Schmid and Trede (1998), Anderson (1996), Davidson

and Duclos (2000), Barrett and Donald (2003) and Linton, Maasoumi and Whang (2005), the present study seeks to combine given models to build a dominance relation in the spirit of Kuosmanen (2004), Roman, Darby-Dowman and Mitra (2006) and Post, Karabati and Arvanitis (2018).

This application of SD seems even more promising than the application to forecast model comparison: optimization enhances the discriminatory power of stochastic orders and, in addition, introduces a greater need for robustness.

Whereas SD comparisons tends to suffer from low discriminatory power due to incomparability of the evaluated forecast models, the proposed methodology can construct forecast combinations which exhibit the desired dominance relation.

Optimization based on a given loss function fuels over-fitting to the data and generally exhibits limited predictive ability out of sample. The over-fitting problem is reduced by requiring improvements for all permissible loss functions, including ones which are more robust against outliers than the standard Gaussian loss function.

The optimal forecast combination weights are found using numerical optimization. A Convex Optimization problem is obtained by applying a linearization of partial moments in the spirit of Rockafellar and Uryasev (2000). Similar linearizations have been used in optimization with standard second-degree SD (SSD) constraints (Dentcheva and Ruszczyński (2003), Kuosmanen (2004) and Roman, Darby-Dowman and Mitra (2006)).

Relevant large-sample properties of the optimal forecast combination can be obtained using functional central limit theorems for the relevant estimator of the joint distribution function, by analogy to the existing asymptotic theory for SD relations in Linton, Maasoumi and Whang (2005), Scaillet and Topaloglou (2010), Linton, Post and Whang (2014) and Post, Karabati and Arvanitis (2018).

The limit theory generalizes results from Post, Karabati and Arvanitis (2018) by allowing for data-dependent slack variables to use results from the econometric literature about set identification. These generalized results can also be applied in portfolio optimization with

standard SSD constraints.

A preliminary analysis reveals that the classes of General Loss functions and Convex Loss functions considered in Jin, Corradi and Swanson (2017, Definition 2.1) have limited discriminatory power because these classes include a range of non-standard, overly permissive loss functions.

For example, for a forecast error distribution supported on $[-1, 1]$, these loss function classes include the loss function $L(E) = (-0.999 - E)$ for $E \in [-1, -0.999]$ and $L(E) = 0$ for $E \in [-0.999, 1]$, which does not penalize errors $E \in [-0.999, 1]$.

It is generally difficult to engineer dominant forecast combinations if improvements are required for such pathological loss functions. This problem is reminiscent of the limited discriminatory power of the First-degree Stochastic Dominance (FSD) order for utility functions.

To enhance power, the focus in the present study is on a class of Symmetric Convex Loss (SCL) functions, which includes the standard Gaussian, Laplacian and Huber loss functions as special cases. The associated stochastic order is tentatively referred to as SCLSD. As will be shown below, this stochastic order is closely related to the standard Second-degree Stochastic Dominance (SSD) order for utility functions.

The flip side of the additional power of the SCL functions is a greater risk of specification error. Several important applications require the relaxation of the symmetry and convexity assumptions or even the use of a 'scoring function' which is more general than a loss function (Gneiting (2011)). Nevertheless, SCL loss functions are appropriate for the common task of the evaluation of the mean or median of a predictive distribution. This task commonly arises in the forecasting of investment returns to portfolios of financial securities and changes to security market indices using predictive regression models.

The proposed methodology is applied to the forecasting of daily changes of the Chicago Board Options Exchange S&P 500 Volatility Index (VIX). The VIX is a leading measure of implied volatility of short-term S&P 500 stock index option prices. The index mirrors the

market price of 'delta-neutral straddles' of S&P500 stock index options; changes in the VIX resemble investment returns to these straddles. Futures and options contracts written upon the VIX provide trading instruments related to market volatility.

The literature about implied volatility predicting starts with Harvey and Whaley (1992) who forecast daily changes of implied volatility of S&P100 index options with a view to trading option positions on the basis of the forecasts. This approach resembles the forecasting of investment returns (of option trading strategies) and should not be confused with using implied volatility estimates for forecasting market volatility (an alternative application area which is reviewed by Poon and Granger (2003)).

The evaluation of volatility forecasts is complicated by the latent nature of true market volatility, which motivates non-standard loss functions such as the QLIKE loss function (Patton (2011)). By contrast, the VIX is an observable measure of expected volatility (rather than realized volatility). The observable nature of the measure allows for using standard predictive regressions for VIX changes, interpreting the resulting forecasts as estimates for the conditional mean of the predictive distribution and evaluating competing specifications using standard loss functions, just as in the more common case of predictive regressions for S&P500 index returns.

The empirical results shows that robust optimized combinations improve significantly upon the out-of-sample forecasting accuracy of simple averaging and unrestricted optimization.

2 Theoretical Concepts

2.1 Preliminaries

A random variable X is forecast using $M \geq 2$ distinct forecast models, generating point

forecasts $\mathbf{Y} := [Y_1 \cdots Y_M]$. The forecasts could be constructed, for example, using predictive regression, analysts forecasts or market prices of securities; the forecasts may also include pre-defined combinations or transformations of basic forecasts. The forecast models are evaluated based on their forecast errors $\mathbf{U} := X\mathbf{1}'_M - \mathbf{Y}$. The joint cumulative distribution function (CDF) of the errors is denoted by $\mathcal{F} : \mathcal{Z}^M \rightarrow [0, 1]$, where $\mathcal{Z} := [a, b]$, $-\infty < a < 0 < b < +\infty$.

The individual forecasts may be biased or imprecise. Combinations of the forecasts are formed to improve the predictive accuracy. This approach can improve precision through diversification of forecast errors. In addition, positive biases of some individual models can offset negative biases of other models. The mixing weights are represented by $\boldsymbol{\lambda} \in \Lambda$, where $\Lambda := \{\boldsymbol{\lambda} \in \mathbb{R}^M : \boldsymbol{\lambda}'\mathbf{1}_M = 1; \boldsymbol{\lambda} \geq \mathbf{0}_M\}$ is the unit simplex.

The errors of a given forecast combination can be expressed as $\mathbf{U}\boldsymbol{\lambda} = X - \mathbf{Y}\boldsymbol{\lambda}$. The marginal CDF is $\mathcal{F}_\lambda(u) = \int_{\{\mathbf{U}:\mathbf{U}\boldsymbol{\lambda} \leq u\}} d\mathcal{F}(\mathbf{U})$.

It is straightforward to generalize the analysis from the unit simplex Λ to a general polytope K . This approach would increase the possibilities for error diversification and bias reduction, but it also tends to increase the sensitivity to sampling variation.

Any polytope can be formulated as the convex hull of its vertices. Therefore, the individual forecasts may be replaced with the most extreme feasible combinations. For example, the base forecasts may include a constant (for example, $Y_1 = 0$) and/or multiples of other forecasts (for example, $Y_1 = cY_2$, $c > 1$), to endogenize the scaling of the forecasts.

An alternative, equivalent approach is obtained by formulating the polytope as the intersection of halfspaces. This formulation however requires a generalization of the proposed notation based on $X - \mathbf{Y}\boldsymbol{\lambda} = \mathbf{U}\boldsymbol{\lambda} - X(\boldsymbol{\lambda}'\mathbf{1}_M - 1)$ instead of $X - \mathbf{Y}\boldsymbol{\lambda} = \mathbf{U}\boldsymbol{\lambda}$. To simplify the notation, the focus here is on the vertex formulation, without loss of generality.

2.2 Stochastic order

Let \mathcal{L} be the class of SCL functions $L : \mathcal{Z} \rightarrow \mathbb{R}_+$ which achieve a minimum at $L(0) = 0$,

increase as the error moves away from zero and obey symmetry: $L(E) = L(|E|)$. This class includes the standard Gaussian, Laplacian and Huber loss functions as special cases.

Instead of pursuing a complete order for a single ‘optimal’ loss function $L \in \mathcal{L}$, the analysis relies on a partial order which considers the entire class of loss functions:

Definition 2.1. *Forecast combination $\lambda \in \Lambda$ stochastically dominates forecast combination $\tau \in \Lambda$ by SCLSD if*

$$\mathbb{E}_{\mathcal{F}} [L(\mathbf{U}\lambda)] \leq \mathbb{E}_{\mathcal{F}} [L(\mathbf{U}\tau)] \quad \forall L \in \mathcal{L}. \quad (1)$$

In other words, SCLSD occurs if the first combination achieves a lower expected loss than the second combination for every permissible loss function. If this dominance relation can be established, then the analyst does not have to choose a specific loss function to rank the two forecasts combinations. It seems particularly comforting that the ranking arises also for loss functions which are robust to outliers.

The class \mathcal{L} is closely related to standard 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 the expectation of $\mathbb{E}_{\mathcal{F}} [U(-|E|)] = -\mathbb{E}_{\mathcal{F}} [L(|E|)]$. SCLSD in terms of forecast error E thus corresponds to SSD in terms of negative absolute forecast error $(-|E|)$.

Every permissible loss function $L \in \mathcal{L}$ is a positive mixture of singularity functions $L_z(x) := (z + |x|)_+$, $z \in \mathcal{Z}^-$, where $(\cdot)_+$ is the positive part and $\mathcal{Z}^- := [\min(a, -b), 0]$. Using this insight, the stochastic order can be reformulated as follows:

Proposition 2.1. *Forecast combination $\lambda \in \Lambda$ stochastically dominates forecast combination $\tau \in \Lambda$ by SCLSD if and only if*

$$\mathbb{E}_{\mathcal{F}} [L_z(\mathbf{U}\lambda)] \leq \mathbb{E}_{\mathcal{F}} [L_z(\mathbf{U}\tau)] \quad \forall z \in \mathcal{Z}^-. \quad (2)$$

For some fixed and given benchmark forecast combination $\boldsymbol{\tau} \in \Lambda$, the set of dominant combinations is given by

$$\Lambda_{\mathcal{F}}^{\succ} := \{\boldsymbol{\lambda} \in \Lambda : \mathbb{E}_{\mathcal{F}} [L_z(\mathbf{U}\boldsymbol{\lambda})] \leq \mathbb{E}_{\mathcal{F}} [L_z(\mathbf{U}\boldsymbol{\tau})] \quad \forall z \in \mathcal{Z}^-\}. \quad (3)$$

This set is non-empty, closed and convex under the maintained assumptions about the joint distribution, loss functions and feasible set Λ .

2.3 Optimal combinations

An optimization problem is proposed to construct a feasible forecast combination which dominates the benchmark. The goal function is the reduction of expected loss for some given loss function, or $G_{\mathcal{F},\boldsymbol{\lambda}} := \mathbb{E}_{\mathcal{F}} [L(\mathbf{U}\boldsymbol{\tau})] - \mathbb{E}_{\mathcal{F}} [L(\mathbf{U}\boldsymbol{\lambda})]$, $L \in \mathcal{L}$. The following optimization problem is proposed:

$$\max_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}^{\succ}} G_{\mathcal{F},\boldsymbol{\lambda}}. \quad (4)$$

If the loss function is strictly convex, the solution will be unique and the optimal combination will be efficient (not dominated by a nonequivalent alternative). In case of local linearity of the loss function, multiple optimal solutions may exist. In this case, a second optimization problem could be solved to detect secondary improvement possibilities and avoid a solution which is inefficient.

Although optimization problem (4) is convex, the number of constraints is uncountable and, furthermore, the dominance constraints ($\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}^{\succ}$) involve the intractable maximum part operator $(\cdot)_+$ in the definition of the singularity function ($L_z(x) := (z + |x|)_+$). Fortu-

nately, an empirical counterpart of the problem can be reformulated as a standard Convex Optimization problem by linearizing the $(\cdot)_+$ operator using additional model variables; see Section 4.4.

3 Numerical Example

A numerical example is developed with two forecast models ($M = 2$) with a simple and known joint error distribution. Analytical solutions can be derived in a straightforward way in this example, in contrast to typical applications. Naturally, the benefits of forecast combination remain limited in the case with only two models compared with a multitude of models.

The random variable X obeys a latent standard uniform distribution. Two independent forecasts are available: Y_1 is a second, independent standard uniform random variable; Y_2 is the mean of *two* of such random variables and thus obey a Bates (1955) distribution. Clearly, both forecasts are unbiased but Y_1 is less precise than Y_2 . For example, the Mean Squared Forecast Error (MSFE) is $\frac{1}{6}$ for Y_1 vs. $\frac{1}{8}$ for Y_2 .

Since both forecasts are based on complementary information, accuracy can be enhanced by combining the two forecasts. The optimal mixing weights are simply $\lambda_1^* = \frac{1}{3}$ and $\lambda_2^* = \frac{2}{3}$, for every loss function $L \in \mathcal{L}$. The optimal forecast follows the Bates distribution for the mean of *three* independent standard uniform variables. This forecast dominates all other mixtures by SCLSD.

For given mixing weights λ_1 and λ_2 , the MSFE equals $(\lambda_1^2 + \frac{1}{2}\lambda_2^2 + 1) \frac{1}{12}$. The optimal forecast reduces the MSFE to $\frac{1}{9} \approx 0.111$, which is the minimum across all mixtures. By contrast, the Equal Weighted Average (EWA) ($\lambda_1 = \lambda_2 = \frac{1}{2}$) achieves an MSFE of $(\frac{11}{8}) (\frac{1}{12}) \approx 0.115$.

Figure 1 illustrates the dominance relation between the optimal mixture on the one hand

and the EWA and best individual forecast (Y_2) on the other hand. The figure shows the expected loss $\mathbb{E}_{\mathcal{F}}[L_z(E)]$ as a function of the threshold level for negative absolute forecast error, $z \in \mathcal{Z}^- = [-1, 0]$. By Proposition 2.1, SCLSD occurs if and only if the optimal mixture reduces expected loss for every threshold level; this condition is clearly satisfied in this example.

[Insert Figure 1 about here]

In this simple example, the unconstrained optimization problem $\max_{\lambda \in \Lambda} G_{\mathcal{F}, \lambda}$ and its constrained counterpart $\max_{\lambda \in \Lambda_{\mathcal{F}}^{\geq}} G_{\mathcal{F}, \lambda}$ have the same solution and, furthermore, the solution does not depend on the choice of the loss function in the objective function. In addition, the joint error distribution \mathcal{F} is known and estimation error for the optimal mixing weights does not play a role.

In more realistic applications, unconstrained optimization for a given loss function does not ensure SD of naive benchmarks. In these cases, explicit SD constraints will ensure reductions in expected loss for every permissible loss function.

Furthermore, in realistic applications, the joint error distribution \mathcal{F} is latent and needs to be estimated, introducing estimation error for the optimal mixing weights. In this case, the SD constraints tend to avoid concentrated solutions and help to improve the robustness and goodness of the solutions. Section 6 studies the effect of sampling variation using a Monte Carlo simulation experiment based on the present numerical example.

4 Empirical Counterparts

The present section introduces an empirical counterpart of optimization problem (4) and

derives a limit theory for the optimal forecasts based on time-series data. The limit theory can be used for statistical inference about the optimal forecasts in the spirit of Diebold and Mariano (1995) and McCracken (2000).

The assumption framework is gradually refined to obtain stronger results. General results obtained under higher-level assumptions are presented in Section 4.1 to 4.4. These assumptions are motivated by more specific lower-level assumptions in Section 4.5. The latter assumptions are similar to those in McCracken (2000) and Jin, Corradi and Swanson (2017) and are consistent with the empirical application in Section 7. Special attention is given to the role of parameter estimation risk for the forecast models.

To derive the limit theory, some additional notation is introduced. 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. $\bar{B}_\lambda(\eta)$ denotes the closed Euclidean ball in \mathbb{R}^M centered at λ with radius equal to $\eta > 0$.

4.1 Empirical optimization problem

The CDF \mathcal{F} is latent and is estimated using a time-series sample of forecasted and realized values of the random variable X . To complicate matters, the forecasts are generally unobservable, since they may depend on unknown parameters.

Using a general formulation, the point forecasts Y_i , $i = 1, \dots, M$, are measurable functions $m_i(\mathbf{Z}_i, \beta_{0_i})$, where $\mathbf{Z}_i \in \mathbb{R}^{d_i}$ is a random vector and $\beta_{0_i} \in \text{Int}B_i$ is the latent parameter vector from parameter space $\Theta_i \subseteq \mathbb{R}^{q_i}$.

The forecasts at time $t \in \{1, 2, \dots, T\}$ are constructed as $\hat{Y}_{i,t} := m_i(Z_{i,t}, \beta_{t_i})$ for realizations $Z_{i,t}$ and parameter estimator β_{t_i} . Given realization x_t , the unobservable error is $\mathbf{U}_t := x_t \mathbf{1}'_M - \mathbf{y}_t$ and the observed error is $\hat{\mathbf{U}}_t := x_t \mathbf{1}'_M - \hat{\mathbf{y}}_t$, where $\mathbf{y}_t := [m_1(Z_{1,t}, \beta_{1_0}) \cdots m_M(Z_{M,t}, \beta_{M_0})]'$ and $\hat{\mathbf{y}}_t := [m_1(Z_{1,t}, \beta_{1_t}) \cdots m_M(Z_{M,t}, \beta_{M_t})]'$.

If m_i is independent of β_{0_i} and/or β_{0_i} is known for all i , then the original error \mathbf{U}_t becomes observable, several of the below assumptions become obsolete, and the derivations become

simpler.

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

$$F_T(\mathbf{u}) := \frac{1}{T} \sum_{t=1}^T \mathbb{I}(\hat{U}_t \leq \mathbf{u}). \quad (5)$$

Section 4.6 includes a discussion about extending the analysis to other discrete estimators for the CDF in the spirit of Post, Karabati and Arvanitis (2018).

To approximate the dominant set $\Lambda_{\mathcal{F}}^{\succeq}$ by an empirical counterpart, it is important to account for boundary problems which may arise in case of binding inequalities; see, for example, Andrews and Soares (2010). These problems are particularly relevant in the present context of optimization with dominance constraints, as the latent optimum $\boldsymbol{\lambda}^*$ will almost always feature binding constraints $\mathbb{E}_{\mathcal{F}}[L_z(\mathbf{u}\boldsymbol{\lambda}^*)] = \mathbb{E}_{\mathcal{F}}[L_z(\mathbf{u}\boldsymbol{\tau})]$ for some $z \in \mathcal{Z}^-$.

These boundary problems are addressed here by introducing random slacks in the right-hand sides of the empirical dominance constraints. Using a stochastic function $c_T : \mathcal{Z}^- \rightarrow \mathbb{R}_+$ to specify the permissible slacks, the following empirical version of $\Lambda_{\mathcal{F}}^{\succeq}$ is employed:

$$\Lambda_{\mathcal{F}_T}^{\succeq}(c_T) := \left\{ \boldsymbol{\lambda} \in \Lambda : \mathbb{E}_{\mathcal{F}_T}[L_z(\mathbf{U}\boldsymbol{\lambda})] \leq \mathbb{E}_{\mathcal{F}_T}[L_z(\mathbf{U}\boldsymbol{\tau})] + c_T(z) \quad \forall z \in \mathcal{Z}^- \right\}, \quad (6)$$

where $\mathbb{E}_{\mathcal{F}_T}[L_z(\mathbf{U}\boldsymbol{\lambda})] = \frac{1}{T} \sum_{t=1}^T L_z(\hat{U}_t\boldsymbol{\lambda})$; the specification of the slack function c_T is discussed in Section 4.3 below.

The empirical counterpart of the goal function is $\mathcal{G}_{\mathcal{F}_T, \boldsymbol{\lambda}} = \mathbb{E}_{\mathcal{F}_T}[L(\mathbf{U}\boldsymbol{\tau})] - \mathbb{E}_{\mathcal{F}_T}[L(\mathbf{U}\boldsymbol{\lambda})]$. Maximizing this goal function subject to the empirical dominance constraints gives the following empirical counterpart of forecast optimization problem (4):

$$\max_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}_T}^{\succeq}(c_T)} \mathbb{E}_{\mathcal{F}_T}[L(\mathbf{U}\boldsymbol{\tau})] - \mathbb{E}_{\mathcal{F}_T}[L(\mathbf{U}\boldsymbol{\lambda})]. \quad (7)$$

4.2 Empirical process

Special attention is given to the SCLSD empirical process $\sqrt{T} (\mathbb{E}_{\mathcal{F}_T} [L_z(\mathbf{U}\boldsymbol{\lambda})] - \mathbb{E}_{\mathcal{F}} [L_z(\mathbf{U}\boldsymbol{\lambda})]) = \frac{1}{\sqrt{T}} \sum_{t=1}^T [L_z(\hat{\mathbf{U}}_t\boldsymbol{\lambda}) - \mathbb{E}_{\mathcal{F}} [L_z(\hat{\mathbf{U}}_t\boldsymbol{\lambda})]]$. The limit theory builds on the higher-level assumption that this empirical process converges weakly to a well-defined, zero-mean Gaussian process:

Assumption 4.1. *As $T \rightarrow \infty$*

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T [L_z(\hat{\mathbf{U}}_t\boldsymbol{\lambda}) - \mathbb{E}_{\mathcal{F}} L_z(\mathbf{U}'_t\boldsymbol{\lambda})] \rightsquigarrow \mathcal{G}(z, \boldsymbol{\lambda}), \text{ in } \ell^\infty(\mathcal{Z}^- \times \boldsymbol{\Lambda}), \quad (8)$$

where \mathcal{G} is a zero-mean Gaussian process with uniformly continuous sample paths in $\ell^\infty(\mathcal{Z}^- \times \boldsymbol{\Lambda})$ and covariance kernel $K_{\mathcal{G}} : (\mathcal{Z}^- \times \boldsymbol{\Lambda})^2 \rightarrow \mathbb{R}$ which is positive definite.

The dependence of $\hat{\mathbf{U}}_t$ on the estimators β_{t_i} , $i = 1, \dots, M$ implies that the asymptotic covariance kernel $K_{\mathcal{G}}$ generally reflects parameter estimation risk in addition to sample variation.

Since every $L \in \mathcal{L}$ can be characterized as $\int_{\mathcal{Z}^-} w(z) L_z dz$ for some positive mixture function $w : \mathcal{Z}^- \rightarrow [0, 1]$, with $\int_{\mathcal{Z}^-} w(z) dz = 1$, the Continuous Mapping Theorem directly implies the following result:

Lemma 4.2. *Under Assumption 4.1, as $T \rightarrow \infty$,*

$$\sqrt{T} (\mathbb{E}_{\mathcal{F}_T} [L(\mathbf{U}\boldsymbol{\lambda})] - \mathbb{E}_{\mathcal{F}} [L(\mathbf{U}\boldsymbol{\lambda})]) \rightsquigarrow \mathcal{G}_L(\boldsymbol{\lambda}), \text{ in } \ell^\infty(\boldsymbol{\Lambda}), \quad (9)$$

where $\mathcal{G}_L(\boldsymbol{\lambda}) := \int_{\mathcal{Z}^-} w(z) \mathcal{G}(\boldsymbol{\lambda}, z) dz$ is a zero-mean Gaussian process with covariance kernel $K_{\mathcal{G}_L}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) := \int_{\mathcal{Z}^- \times \mathcal{Z}^-} w(z) w(z^*) K_{\mathcal{G}}((\boldsymbol{\lambda}_1, z), (\boldsymbol{\lambda}_2, z^*)) dz dz^*$.

Assumption 4.1 is helpful to determine whether $\Lambda_{\mathcal{F}_T}^{\succeq}(c_T)$ approximates $\Lambda_{\mathcal{F}}^{\succeq}$. Lemma 4.2 implies that the empirical objective function $\mathcal{G}_{\mathcal{F}_T, \boldsymbol{\lambda}}$ converges in probability to its population

equivalent $\mathcal{G}_{\mathcal{F},\lambda}$ uniformly in λ .

4.3 Consistency properties

Consistency properties are derived for (i) the empirical dominance classification, (ii) the optimal value of the goal function and (iii) the optimal values of the mixing weights.

In what follows, λ_T denotes an optimal forecast combination, or a solution to the empirical optimization problem (7).

The following proposition establishes that false dominance classifications asymptotically almost never occur:

Proposition 4.3. *If Assumption 4.1 holds and $\sup_{z \in \mathcal{Z}} c_T(z) \rightarrow 0$, then as $T \rightarrow \infty$, $\lambda_T \in \Lambda_{\mathcal{F}}^{\succ}$, with probability converging to one.*

While this result ensures that λ_T converges to a feasible solution of latent optimization problem (4), it does not suffice to guarantee that λ_T converges to an optimal solution. To achieve stronger results, several additional assumptions and concepts are introduced.

To deal with the aforementioned boundary problem, the assumptions on the slack process are tightened: $c_T(z)$ is now assumed to be strictly positive and to converge to zero in probability with rate slower than \sqrt{T} as $T \rightarrow \infty$. A deterministic example is $c_T = f(T)T^{-0.5}$ for all z , where f is a slowly varying function at infinity, for example, the logarithmic function. The econometric literature provides further guidance for the optimal specification of slacks; see, for example, Andrews and Soares (2010) and the references therein.

To establish the asymptotic approximation of $\Lambda_{\mathcal{F}}^{\succ}$ by $\Lambda_{\mathcal{F}_T}^{\succ}(c_T)$, a stochastic version of Painleve-Kuratowski set convergence (see Molchanov (2006, Appendix B)) is used. A sequence of non-empty closed subsets $\Lambda_T \subseteq \Lambda$ is said to Painleve-Kuratowski converge to a closed subset $\Lambda^* \subseteq \Lambda$, if and only if any $\lambda \in \Lambda^*$ is simultaneously a cluster point of some sequence λ_T and a limit point in probability of some sequence λ_T^* , with $\lambda_T, \lambda_T^* \in \Lambda_T$ for all T . More generally, if Λ_T and/or Λ^* are random closed subsets of Λ (in the sense of Molchanov (2006, Chapter 1)), then Λ_T Painleve-Kuratowski converges to Λ^* in probability, if and only

if the aforementioned limit and cluster point properties hold with probability converging to one.

In case of multiple optimal solutions to problem (4), several subsequences of $\boldsymbol{\lambda}_T$ might approximate different solutions. In order to ensure a unique solution, strict convexity is assumed for the loss function L in the goal function as well as linear independence of the errors of the various forecast models (\mathbf{U}).

The following theorem establishes various additional consistency properties using the aforementioned assumptions:

Theorem 4.4. *Suppose that $\sup_{z \in \mathcal{Z}^-} c_T(z) \rightarrow 0$ and $\sqrt{T} \inf_{z \in \mathcal{Z}^-} c_T(z) \rightarrow \infty$ in probability. Then, under Assumption 4.1, as $T \rightarrow \infty$, a) $\Lambda_{\mathcal{F}_T}^{\succeq}(c_T)$ converges in probability to $\Lambda_{\mathcal{F}}^{\succeq}$ in the Painleve-Kuratowski sense and, consequently, b) $\mathbb{E}_{\mathcal{F}_T}[L(\mathbf{U}\boldsymbol{\tau})] - \mathbb{E}_{\mathcal{F}_T}[L(\mathbf{U}\boldsymbol{\lambda}_T)]$ converges in probability to $\max_{\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}^{\succeq}} G_{\mathcal{F},\boldsymbol{\lambda}}$. If moreover L is strictly convex and the covariance matrix of \mathbf{U} is positive definite then, c) $\boldsymbol{\lambda}_T \rightsquigarrow \boldsymbol{\lambda}_{\mathcal{F}}^*$ where $\boldsymbol{\lambda}_{\mathcal{F}}^*$ is the unique solution to (4).*

The assumption of strict convexity applies for the Gaussian loss function which underlies the MSFE goal function which is used in the application section. However, it does not hold for the Laplacian loss function which underlies the Mean Absolute Forecast Error (MAFE). In the latter case, multiple optimal forecast combinations may exist and point convergence cannot be established. Nevertheless, it can still be established that the optimal solution features asymptotically vanishing decision errors for the dominance classification and the optimal value of the goal function; Proposition 4.3 and Theorem 4.4b do not require strict convexity.

4.4 Rates and limiting distributions

The limit distributions of the optimal forecast combinations are derived using the above results, a local quadratic approximation for $\mathbb{E}_{\mathcal{F}_T}[L(\cdot)]$ in a neighborhood of $\boldsymbol{\lambda}_{\mathcal{F}}^*$, and the limiting behavior of the coefficients of the approximation. The assumption framework is further tightened using the following assumptions:

Assumption 4.5. *The following conditions are satisfied:*

i. *For some $\eta > 0$ and any $\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}^{\succ} \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta)$,*

$$\begin{aligned} & T \left(\mathbb{E}_{\mathcal{F}_T} \left[L \left(\mathbf{U} \left(\boldsymbol{\lambda}_{\mathcal{F}}^* + \frac{1}{\sqrt{T}} \mathbf{u} \right) \right) \right] - \mathbb{E}_{\mathcal{F}_T} [L(\mathbf{U} \boldsymbol{\lambda}_{\mathcal{F}}^*)] \right) \\ &= \sqrt{T} \mathbb{E}_{\mathcal{F}_T} [s_t(\boldsymbol{\lambda}_{\mathcal{F}}^*)]^T \mathbf{u} + \frac{1}{2} \mathbf{u}^T \mathbb{E}_{\mathcal{F}_T} [\mathcal{H}_t(\boldsymbol{\mu}_{\mathcal{F}}^*)] \mathbf{u}, \end{aligned} \quad (10)$$

where $s_t(\cdot)$ is a random $M \times 1$ vector function and $\mathcal{H}_t(\cdot)$ a random $M \times M$ matrix function defined on $\Lambda_{\mathcal{F}}^{\succ} \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta)$, $\mathbf{u} := \sqrt{T}(\boldsymbol{\lambda} - \boldsymbol{\lambda}_{\mathcal{F}}^*)$ and $\boldsymbol{\mu}_{\mathcal{F}}^*$ is some element of $\Lambda_{\mathcal{F}}^{\succ} \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta)$ lying on the ray that connects $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}_{\mathcal{F}}^*$.

ii. *As $T \rightarrow \infty$, for some random vectors $s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*)$, $S_{\boldsymbol{\lambda}_{\mathcal{F}}^*}$,*

$$\sqrt{T} (\mathbb{E}_{\mathcal{F}_T} [s_t(\boldsymbol{\lambda}_{\mathcal{F}}^*)] - \mathbb{E}_{\mathcal{F}} [s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*)]) \rightsquigarrow S_{\boldsymbol{\lambda}_{\mathcal{F}}^*} \sim N(\mathbf{0}, V_{\boldsymbol{\lambda}_{\mathcal{F}}^*}), \quad (11)$$

with $V_{\boldsymbol{\lambda}_{\mathcal{F}}^*}$ an $M \times M$ matrix, and for any $\lambda_T \rightarrow \boldsymbol{\lambda}_{\mathcal{F}}^*$,

$$\mathbb{E}_{\mathcal{F}_T} [\mathcal{H}_t(\lambda_T)] \rightsquigarrow \mathcal{H}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}, \quad (12)$$

with $\mathcal{H}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}$ a positive definite $M \times M$ matrix.

iii. *If $\mathbb{E}_{\mathcal{F}} [s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*)] \neq \mathbf{0}_M$, and moreover, $\sqrt{T}(\boldsymbol{\lambda}_T - \boldsymbol{\lambda}_{\mathcal{F}}^*) = O_p(1)$, then $\mathbb{E}_{\mathcal{F}} [L(\mathbf{U} \boldsymbol{\lambda})]$ is continuously differentiable in a neighborhood of $\boldsymbol{\lambda}_{\mathcal{F}}^*$, and its derivative at $\boldsymbol{\lambda}_{\mathcal{F}}^*$ equals $\mathbb{E}_{\mathcal{F}} [s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*)]$.*

Assumption 4.5.(i) can be derived, assuming $\boldsymbol{\lambda}_{\mathcal{F}}^* \neq \boldsymbol{\tau}$, from continuous second-order differentiability of L on $[a, b]$ and the relevant first-order Taylor expansion of L with second-order remainders obtained via the Mean Value Theorem, as in Andrews (1999). In the

present case, $s_t(\lambda) = \frac{dL}{dx}(\hat{U}_t \boldsymbol{\lambda}) \hat{U}_t$ and $H_t(\lambda) = \frac{d^2 L}{dx^2}(\hat{U}_t \boldsymbol{\lambda}) \hat{U}_t' \hat{U}_t$, and under stationarity, $s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*) = \frac{dL}{dx}(\mathbf{U}_0 \boldsymbol{\lambda}_{\mathcal{F}}^*) \mathbf{U}_0$, where the derivatives involved may be one-sided and, furthermore, (10) holds for any $\eta > 0$, given the convexity of $\Lambda_{\mathcal{F}}^{\succeq}$.

Assumption 4.5.(ii) can be derived via some relevant Central Limit Theorem; see Section 4.5 below.

The tightness in Assumption 4.5.(iii) is derivable via results like Theorem 5.52 of van der Vaart (2000) adjusted to dependent data. The identification of the derivative would then follow by a dominated convergence argument. The case $\mathbb{E}_{\mathcal{F}}[s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*)] \neq \mathbf{0}_M$ holds for example for the Gaussian loss function.

Equipped with these additional assumptions, the limit distribution of the optimal mixing weights can now be derived.

Theorem 4.6. *Suppose that Assumptions 4.1 and 4.5 hold, the function $\boldsymbol{\lambda} \rightarrow \mathbb{E}_{\mathcal{F}}[L(\mathbf{U}\boldsymbol{\lambda})]$ is strictly convex, and $\boldsymbol{\lambda}_T$ solves (7).*

If $\mathbb{E}_{\mathcal{F}}[s^(\boldsymbol{\lambda}_{\mathcal{F}}^*)] = \mathbf{0}_M$ then as $T \rightarrow \infty$,*

$$\sqrt{T}(\boldsymbol{\lambda}_T - \boldsymbol{\lambda}_{\mathcal{F}}^*) \rightsquigarrow \boldsymbol{\lambda}_{\infty}^*, \quad (13)$$

where

$$\boldsymbol{\lambda}_{\infty}^* = \arg \min_{\mathbf{u} \in \Lambda_{\infty}^{\succeq}} \left\| \mathcal{H}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}^{\frac{1}{2}} \mathbf{u} + \mathcal{H}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}^{-\frac{1}{2}} S_{\boldsymbol{\lambda}_{\mathcal{F}}^*} \right\|,$$

$\Lambda_{\infty}^{\succeq}$ is the non-empty closed and convex Painleve-Kuratowski limit of $\sqrt{T}(\Lambda_{\mathcal{F}}^{\succeq} \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta) - \boldsymbol{\lambda}_{\mathcal{F}}^*)$, and $\|\cdot\|$ denotes the Euclidean norm.

If $\mathbb{E}_{\mathcal{F}}[s^(\boldsymbol{\lambda}_{\mathcal{F}}^*)] \neq \mathbf{0}_M$ and $\Lambda_{\mathcal{F}}^{\succeq} \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta) - \boldsymbol{\lambda}_{\mathcal{F}}^*$ is locally at zero equal to a cone, then as $T \rightarrow \infty$, (13) holds with*

$$\boldsymbol{\lambda}_{\infty}^* = \mathbf{0}_M.$$

Whenever $\mathbb{E}_{\mathcal{F}}[s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*)] = \mathbf{0}_M$, the limit distribution is not degenerate. The limit distribution is essentially obtained by the projection via the norm $\sqrt{(\cdot)^T H_t(\boldsymbol{\lambda}_{\mathcal{F}}^*) (\cdot)}$ of the zero-mean normal random vector $-H_{\boldsymbol{\lambda}_{\mathcal{F}}^*}^{-1} S_{\boldsymbol{\lambda}_{\mathcal{F}}^*}$ to the convex $\Lambda_{\infty}^{\succeq}$. Standard \sqrt{T} rates are obtained. The

limiting distribution depends on system of inequalities (2) and thereby on the benchmark $\boldsymbol{\tau}$, because the asymptotic variance depends on $\boldsymbol{\lambda}_{\mathcal{F}}^*$ and the limiting $\Lambda_{\infty}^{\succeq}$ depends on $\boldsymbol{\lambda}_{\mathcal{F}}^*$ and $\Lambda_{\mathcal{F}}^{\succeq}$.

Since the elements of $\Lambda_{\mathcal{F}}^{\succeq}$ must satisfy the simplex constraints, $\Lambda_{\infty}^{\succeq} \neq \mathbb{R}^M$. If $\boldsymbol{\lambda}_{\mathcal{F}}^*$ is an interior point of $\Lambda_{\mathcal{F}}^{\succeq}$, then $\Lambda_{\infty}^{\succeq}$ is a hyperplane and $\boldsymbol{\lambda}_{\infty}^*$ has a singular distribution; if $\boldsymbol{\lambda}_{\mathcal{F}}^*$ is a boundary point, then $\boldsymbol{\lambda}_{\infty}^*$ will also have a singular distribution but the distribution will be more concentrated.

Whenever $\mathbb{E}_{\mathcal{F}}[s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*)] \neq \mathbf{0}_M$, the assumption framework ensures that the limiting criterion is asymptotically dominated by $-\sqrt{T}\mathcal{H}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}^{-1}\mathbb{E}_{\mathcal{F}}[s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*)]$. This term is then asymptotically projected to the zero element of $\Lambda_{\infty}^{\succeq}$, due to its position in the normal cone of $\Lambda_{\mathcal{F}}^{\succeq}$, leading to degeneracy. The local equality to a cone property for $\Lambda_{\mathcal{F}}^{\succeq} \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta) - \boldsymbol{\lambda}_{\mathcal{F}}^*$ follows directly from its convexity, if $M = 2$. For more general cases, see, for example, Andrews (1999, Section 4.2).

4.5 Refinements based on lower-level assumptions

The general results obtained in Sections 4.1 to 4.4 are now motivated by, and further developed for, specific lower-level assumptions. These assumptions are similar to those in McCracken (2000) and Jin, Corradi and Swanson (2017) and are consistent with the empirical application in Section 7.

Assumption 4.1'. *Suppose that the following conditions hold:*

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

- iii. The vector process $\mathbf{Z}_t := \left[x_t, (Z_{i,t}, h_{i,t})_{i=1, \dots, M} \right]_{t \in \mathbb{Z}}$ is strictly stationary and strongly mixing with mixing coefficients $(\alpha_k)_{k \in \mathbb{N}}$ that satisfy $\alpha_k = O(k^{-r})$ for $r > Q$ and $Q := \left(1 + 2 \left(M + \prod_{i=1}^M q_i\right)\right) \left(2 + M + \prod_{i=1}^M q_i\right)$.
- iv. For some $\eta > 0$, such that for $\beta := (\beta_1, \dots, \beta_M)$ restricted to $\bar{B}_{\beta_0}(\eta) \subset \mathbb{R}^{\prod_{i=1}^M q_i}$, and $\beta_0 = (\beta_{1_0}, \dots, \beta_{M_0})$, the function $\beta \rightarrow K(\mathbf{Z}_0, \beta) := x_0 \mathbf{1}'_M - [m_1(Z_{1,0}, \beta_1) \cdots m_M(Z_{M,0}, \beta_M)]$ is almost surely Lipschitz continuous with respect to β , with Lipschitz coefficient $l(\mathbf{Z}_0)$, that satisfies $\mathbb{E}[l^2(\mathbf{Z}_0)] < +\infty$.
- v. The function $(z, \boldsymbol{\lambda}, \beta) \rightarrow G(z, \boldsymbol{\lambda}, \beta) := \mathbb{E}_{\mathcal{F}}[L_z(K(\mathbf{Z}_0, \beta) \boldsymbol{\lambda})]$ is continuously differentiable with respect to β on $\bar{B}_{\beta_0}(\eta)$, for all $(z, \boldsymbol{\lambda}) \in \mathcal{Z}^- \times \boldsymbol{\Lambda}$ and $\sup_{\mathcal{Z}^- \times \boldsymbol{\Lambda} \times \bar{B}_{\beta_0}(\eta)} \|D_{\beta} G(z, \boldsymbol{\lambda}, \beta)\| < \infty$.

Assumptions (i)-(ii) are satisfied when the estimators of the unobserved parameters are evaluated at a rolling window of R_T observations and the window size R_T is of the same asymptotic order as T . Assumption (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, under the appropriate conditions.

Assumption (iii) holds for common strictly stationary versions of multivariate ARMA, GARCH and stochastic volatility models and measurable transformations. Assumption (iv) holds for predictive regressions, in which case m_i is bilinear in $(Z_{i,0}, \beta_i)$, as long as the regressors have second moments. Whenever \mathcal{F} has a density and m_i is almost surely continuously differentiable with respect to β_i , Assumption (v) can be verified by arguments similar to the ones in Section 5 of Kim and Pollard (1990).

Lemma 4.7. *Assumption 4.1' implies Assumption 4.1 with*

$$\begin{aligned}
K_G((z_1, \boldsymbol{\lambda}_1), (z_2, \boldsymbol{\lambda}_2)) &= \sum_{i=0}^{\infty} \kappa_i \text{Cov}(L_{z_1}(\mathbf{U}_0 \boldsymbol{\lambda}_1), L_{z_2}(\mathbf{U}_i \boldsymbol{\lambda}_2)) + \\
&+ \varrho \sum_{i=0}^{\infty} \kappa_i \text{Cov}(L_{z_1}(\mathbf{U}_0 \boldsymbol{\lambda}_1), D_{\beta} G(z_2, \boldsymbol{\lambda}_2, \beta_0) \mathbf{H} \mathbf{h}_i) \\
&+ \varrho \sum_{i=0}^{\infty} \kappa_i \text{Cov}(L_{z_2}(\mathbf{U}_0 \boldsymbol{\lambda}_2), D_{\beta} G(z_1, \boldsymbol{\lambda}_1, \beta_0) \mathbf{H} \mathbf{h}_i) \\
&+ \varrho_{\star} D_{\beta} G(z_1, \boldsymbol{\lambda}_1, \beta_0) \mathbf{H} V_h \mathbf{H}' D_{\beta} G(z_2, \boldsymbol{\lambda}_2, \beta_0)'
\end{aligned} \tag{14}$$

for $z_1, z_2 \in \mathcal{Z}^-$, $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \boldsymbol{\Lambda}$, and $\kappa_i = \begin{cases} 1, & i = 0 \\ 2, & i > 0 \end{cases}$. Here, $D_{\beta} G := \frac{\partial G}{\partial \beta'}$, \mathbf{H} is the $\prod_{i=1}^M q_i \times$

$\prod_{i=1}^M q_i$ block diagonal matrix $\text{diag}_{1 \leq i \leq \prod_{i=1}^M q_i} (H_{0i})$, $\mathbf{h}_t := (h_{i,t})'_{i=1, \dots, M}$, $V_h := \sum_{i=0}^{\infty} \kappa_i \mathbb{E}[\mathbf{h}_0 \mathbf{h}_i']$,
and $\varrho = \begin{cases} 1 - \frac{\gamma}{2}, & \gamma < 1 \\ \frac{1}{2\gamma}, & \gamma \in [1, +\infty] \end{cases}$, $\varrho_{\star} = \begin{cases} 1 - \frac{\gamma}{3}, & \gamma < 1 \\ \frac{1}{\gamma} - \frac{1}{3\gamma^2}, & \gamma \in [1, +\infty] \end{cases}$.

The covariance kernel in (14) reflects the sample error variation through the first term, the estimated parameters error variation through the last term and the covariation between the two errors through the remaining terms. When $D_{\beta} G(z, \boldsymbol{\lambda}, \beta_0) = \mathbf{0}_{1 \times \prod_{i=1}^M q_i}$ for all $z, \boldsymbol{\lambda}$, and/or $\gamma = \infty$, any (co-) variation due to the estimated parameters error disappears.

Assumption 4.5'. *Suppose that the following conditions are satisfied:*

- i. For any $\mathbf{u} \in \mathbb{R}^M$, the function $\mathbb{R}^M \ni \boldsymbol{\lambda} \rightarrow L(\mathbf{u}' \boldsymbol{\lambda})$ is continuously second-order differentiable.
- ii. Assumption 4.1'.(i)-(iii) holds. If $\mathbb{E}_{\mathcal{F}} \left[\frac{dL}{dx}(\mathbf{U}_0 \boldsymbol{\lambda}_{\mathcal{F}}^*) \mathbf{U}_0 \right] \neq \mathbf{0}_M$, then the mixing part of Assumption 4.1.(iii) is strengthened to absolute regularity, where the analogous mixing coefficients have the same rates.
- iii. The function $\beta \rightarrow \frac{dL}{dx}(K(\mathbf{Z}_0, \beta) \boldsymbol{\lambda}_{\mathcal{F}}^*) K(\mathbf{Z}_0, \beta)$ is almost surely Lipschitz continuous on $\bar{B}_{\beta_0}(\eta)$, with Lipschitz coefficient $k(\mathbf{Z}_0)$, that satisfies $\mathbb{E}[k^2(\mathbf{Z}_0)] < +\infty$, where K, β

and $\bar{B}_{\beta_0}(\eta)$ are as in Assumption 4.1'.

iv. The function $\beta \rightarrow Q_{\lambda_{\mathcal{F}}^*}(\beta) := \mathbb{E} \left[\frac{dL}{dx} (K(\mathbf{Z}_0, \beta) \lambda_{\mathcal{F}}^*) K(\mathbf{Z}_0, \beta) \right]$ is continuously differentiable on $\bar{B}_{\beta_0}(\eta)$.

v. For some $\eta' > 0$, $\mathbb{E}_{\mathcal{F}} \left[\sup_{\lambda \in \bar{B}_{\lambda_{\mathcal{F}}^*}(\eta'), \beta \in \bar{B}_{\beta_0}(\eta)} \left| \frac{d^2 L(K(\mathbf{Z}_0, \beta) \lambda)}{dx^2} \right| \|K(\mathbf{Z}_0, \beta)' K(\mathbf{Z}_0, \beta)\| \right] < +\infty$.

vi. The matrix $\mathbf{U}'_0 \mathbf{U}_0$ is positive definite with positive probability.

Given condition (i), conditions (iii)-(v) follow, for example, if the random element \mathbf{Z}_0 has sufficiently high moments, in the case of predictive regressions and if the loss function is dominated by a polynomial for large values of its argument. The absolute regularity condition in (ii) is also satisfied for several popular models; see, for example, Section 4 of Mikosch and Straumann (2006) for the GARCH model. Similarly to the third part of Theorem 4.4, Assumption (vi) would follow if the elements of \mathbf{U}_0 are linearly independent.

Lemma 4.8. *Assumption 4.5' implies Assumption 4.5 with $s^*(\lambda_{\mathcal{F}}^*) = \frac{dL}{dx}(\mathbf{U}_0 \lambda_{\mathcal{F}}^*) \mathbf{U}_0$,*

$$\begin{aligned} V_{\lambda_{\mathcal{F}}^*} &= \sum_{i=0}^{\infty} \kappa_i \text{Cov}_{\mathcal{F}} \left(\frac{dL(\mathbf{U}_0 \lambda_{\mathcal{F}}^*)}{dx} \mathbf{U}_0, \frac{dL(\mathbf{U}_i \lambda_{\mathcal{F}}^*)}{dx} \mathbf{U}_i \right) + \\ &+ 2\varrho \sum_{i=0}^{\infty} \kappa_i \text{Cov} \left(\frac{dL(\mathbf{U}_0 \lambda_{\mathcal{F}}^*)}{dx} \mathbf{U}_0, D_{\beta} Q_{\lambda_{\mathcal{F}}^*}(\beta_0) \mathbf{H} \mathbf{h}_i \right), \\ &+ \varrho_{\star} D_{\beta} Q_{\lambda_{\mathcal{F}}^*}(\beta_0) \mathbf{H} V_h \mathbf{H}' D_{\beta} Q_{\lambda_{\mathcal{F}}^*}(\beta_0)' \end{aligned}$$

and

$$\mathcal{H}_{\lambda_{\mathcal{F}}^*} = \mathbb{E}_{\mathcal{F}} \left[\frac{d^2 L(\mathbf{U}_0 \lambda_{\mathcal{F}}^*)}{dx^2} \mathbf{U}'_0 \mathbf{U}_0 \right],$$

where $D_{\beta} Q_{\lambda_{\mathcal{F}}^*} := \frac{\partial Q_{\lambda_{\mathcal{F}}^*}}{\partial \beta'}$, \mathbf{H} is the $\prod_{i=1}^M q_i \times \prod_{i=1}^M q_i$ block diagonal matrix $\text{diag}_{1 \leq i \leq \prod_{i=1}^M q_i} (H_{0i})$,

$$\mathbf{h}_t := (h_{i,t})'_{i=1, \dots, M}, V_h := \sum_{i=0}^{\infty} \kappa_i \mathbb{E} [\mathbf{h}_0 \mathbf{h}'_i], \text{ and } \varrho = \begin{cases} 1 - \frac{\gamma}{2}, & \gamma < 1 \\ \frac{1}{2\gamma}, & \gamma \in [1, +\infty] \end{cases}, \varrho_{\star} = \begin{cases} 1 - \frac{\gamma}{3}, & \gamma < 1 \\ \frac{1}{\gamma} - \frac{1}{3\gamma^2}, & \gamma \in [1, +\infty] \end{cases}.$$

The limiting variance $V_{\lambda_{\mathcal{F}}^*}$ again reflects the sample variation, the estimated parameter error variation and the covariation between the two error terms. Similar to the results of McCracken (2000), the estimated parameter error affects the variance unless $D_{\beta}Q_{\lambda_{\mathcal{F}}^*}(\beta_0) = 0_{1 \times \prod_{i=1}^M q_i}$ and/or $\gamma = \infty$. The limiting Hessian is in any case unaffected by parameter estimation error due to the $\frac{1}{T}$ scaling of the empirical Hessian.

Under Assumptions 4.1' and 4.5', the convergence in Lemma 4.2 and Theorem 4.6 is joint. Given consistency, the Continuous Mapping Theorem, a usual Taylor expansion and the fact that the term $\mathbb{E}_{\mathcal{F}} [s_0(\lambda_{\mathcal{F}}^*)]' \lambda_{\infty}^*$ equals zero in both cases of Theorem 4.6, the following corollary is obtained.

Corollary 4.9. *Under Assumptions 4.1' and 4.5' and the premises of Theorem 4.6, as $T \rightarrow \infty$,*

$$\sqrt{T} (\mathbb{E}_{\mathcal{F}_T} [L(\mathbf{U}\lambda_T)] - \mathbb{E}_{\mathcal{F}} [L(\mathbf{U}\lambda_{\mathcal{F}}^*)]) \rightsquigarrow \mathcal{G}_L(\lambda_{\mathcal{F}}^*). \quad (15)$$

This result is useful for statistical inference about the optimal forecast combination in the spirit of Diebold and Mariano (1995, Section 1.1) or, if parameter estimation error affects the limiting variance, McCracken (2000, Thm 2.3.2).

4.6 Extensions

In the framework of Assumption 4.1', the above results can be easily extended to allow recursive and/or fixed sampling schemes for the construction of the estimators β_{t_i} for some $i \in \{1, \dots, M\}$. Using the results of West and McCracken (1998) and McCracken (2000), and extending Assumption 4.1' accordingly, Lemmata 4.7 and 4.8 would continue to hold, featuring however more complicated expressions for K_G and $V_{\lambda_{\mathcal{F}}^*}$. Furthermore, the extension of the results to general prediction horizons is straightforward.

Another possible extension involves the use of the implied CDF (ICDF) estimators of Gen-

eralized Methods of Moments (Back and Brown (1993) and Generalized Empirical Likelihood (Qin and Lawless (1994)), which can account for side information and dynamic patterns that are not captured by the ECDF.

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 bias can sometimes be determined based on accounting conventions such as 'conservatism' or modeling assumptions such as 'risk neutrality'. In addition, side information about the number and noise level of observations used in predictive regressions could be used to impose a prior ranking for the individual forecasts models.

To account for dynamic patterns, the analysis could build on the Blockwise Empirical Likelihood (BEL) methodology (Kitamura (1997)), extending earlier work by Post, Karabati and Arvanitis (2018) in the context of portfolio optimization with standard SSD constraints. Assumption 4.1' would have to be extended to address the side information moment conditions and the asymptotic relation between the block size, R and T . Relevant generalizations of Lemmata 4.7 and 4.8 would reveal asymptotic efficiency gains in results like Corollary 4.9, compared to the ECDF case stemming from the use of side information and blocking schemes. These efficiency gains would however be mitigated by the presence of parameter estimation error.

The analysis thus far has worked with the simple slack function $c_T = f(T)T^{-0.5}$ for all z , where f is a positive function slowly diverging to infinity. This specification is sample independent and threshold independent specification and may be suboptimal if one is interested in the quality of the approximation of $\Lambda_{\mathcal{F}}^{\succ}$ by $\Lambda_{\mathcal{F}_T}^{\succ}(c_T)$ in small samples. In such cases, the problem of optimal choice of slacks especially in the presence of parameter estimation error merits further research.

5 Numerical Optimization

5.1 Optimization problem

The empirical application will focus on the case with constant slacks: $c_T(z) = c_T \forall z \in \mathcal{Z}^-$. In this case, the empirical dominance conditions in (6) can be reduced to the following finite system

$$\mathbb{E}_{\mathcal{F}_T} [L_z(\mathbf{U}\boldsymbol{\lambda})] \leq \mathbb{E}_{\mathcal{F}_T} [L_z(\mathbf{U}\boldsymbol{\tau})] + c_T \quad \forall z \in \mathcal{Z}_T^-. \quad (16)$$

The partial moments are evaluated only at the elements of \mathcal{Z}_T^- rather than the entire support \mathcal{Z}^- . This reduction is allowed because $\mathbb{E}_{\mathcal{F}_T} [L_z(\mathbf{U}\boldsymbol{\tau})]$ by construction is a piece-wise linear, convex function of the threshold $z \in \mathcal{Z}^-$ with kinks only at the elements of \mathcal{Z}_T^- ; consequently, if the constraints are satisfied for all $z \in \mathcal{Z}_T^-$, then they are also satisfied for all $z \in \mathcal{Z}^-$. A similar result arises in Bawa, Bodurtha, Suri and Rao (1985, Section IC.1) for standard SD criteria for pairwise and multiple comparison.

Let \mathbf{z} a $(T \times 1)$ vector with the ranked elements of \mathcal{Z}_T^- and $\boldsymbol{\sigma}$ a $(T \times 1)$ vector of corresponding values of $(\mathbb{E}_{\mathcal{F}_T} [L_z(\mathbf{U}\boldsymbol{\tau})] + c_T)$, $z \in \mathcal{Z}_T^-$.

Using this notation, the empirical dominance conditions (16) can be rewritten as follows:

$$\mathbf{p}' (\mathbf{1}_T \mathbf{z}' + |\mathbf{E}\boldsymbol{\lambda}| \mathbf{1}'_T)_+ \leq \boldsymbol{\sigma}'. \quad (17)$$

A convenient linearization is obtained in the spirit of Rockafellar and Uryasev (2000):

$$\begin{aligned}
\mathbf{p}'\Theta &\leq \boldsymbol{\sigma}' & (18) \\
-\Theta - \mathbf{E}\lambda\mathbf{1}'_T &\leq -\mathbf{1}_T\mathbf{z}' \\
-\Theta + \mathbf{E}\lambda\mathbf{1}'_T &\leq -\mathbf{1}_T\mathbf{z}' \\
\Theta &\geq \mathbf{0}_{T\times T}
\end{aligned}$$

Here, Θ are additional model variables which capture the element-wise positive parts $(\mathbf{1}_T\mathbf{z}' + |\mathbf{E}\lambda\mathbf{1}'_T|)_+$ if system (17) is solvable.

The empirical counterpart of the goal function, can be formulated as $\mathcal{G}_{\mathcal{F}_T, \lambda} = \mathbf{p}'\mathbf{L}(\mathbf{E}\lambda)$, using the vector-valued function $\mathbf{L}(\boldsymbol{\varepsilon}) := [L(\varepsilon_1) \cdots L(\varepsilon_T)]'$.

The full optimization problem follows:

$$\begin{aligned}
\max \mathbb{E}_{\mathcal{F}_T} [L_z(\mathbf{U}\boldsymbol{\tau})] - \mathbf{p}'\mathbf{L}(\mathbf{E}\lambda) & & (19) \\
\text{s.t. } \mathbf{p}'\Theta &\leq \boldsymbol{\sigma}' \\
-\Theta - \mathbf{E}\lambda\mathbf{1}'_T &\leq -\mathbf{1}_T\mathbf{z}' \\
-\Theta + \mathbf{E}\lambda\mathbf{1}'_T &\leq -\mathbf{1}_T\mathbf{z}' \\
\Theta &\geq \mathbf{0}_{T\times T} \\
\lambda &\in \Lambda
\end{aligned}$$

Since $T \gg M$, the number of variables and constraints is $\mathcal{O}(T^2)$ and increases at a quadratic rate with the number of time-series observations. However, the problem is perfectly tractable for samples of hundreds of observations with standard computer hardware and software. For very large samples of high-frequency data, high-performance platforms and

specialized solver software are recommended.

5.2 Special cases

For important special cases of the goal function, the Convex Optimization problem reduces to a standard Linear Programming or Convex Quadratic Programming problem.

The goal function based on loss function $L_1(U) = |U|$ is the MAFE, $\mathcal{G}_{\mathcal{F},\lambda} = \mathbb{E}_{\mathcal{F}} [L(\mathbf{U}\boldsymbol{\tau})] - \mathbb{E}_{\mathcal{F}} [|\mathbf{U}\boldsymbol{\lambda}|]$. A simple way to include this goal function assumes that the zero vector $\boldsymbol{\varepsilon}_T = \mathbf{0}'_M$ has been included in error data matrix \mathbf{E} (which is inconsequential for system (18)). Using \mathbf{e}_i for the i -th unit vector of proper dimensions, so that $(\boldsymbol{\Theta}\mathbf{e}_T)$ is the last column of $\boldsymbol{\Theta}$, the goal function can be linearized as follows:

$$\mathbb{E}_{\mathcal{F}_T} [|\mathbf{U}\boldsymbol{\lambda}|] = \mathbf{p}'\mathbf{L}_1(\mathbf{E}\boldsymbol{\lambda}) = \mathbf{p}'\boldsymbol{\Theta}\mathbf{e}_T, \quad (20)$$

Similarly, a convex quadratic goal function is obtained when the goal is the MSFE $\mathcal{G}_{\mathcal{F},\lambda} = \mathbb{E}_{\mathcal{F}} [(\mathbf{U}\boldsymbol{\lambda})^2]$, which is based on $L_2(U) = U^2$:

$$\mathbb{E}_{\mathcal{F}_T} [(\mathbf{U}\boldsymbol{\lambda})^2] = \mathbf{p}'\mathbf{L}_2(\mathbf{E}\boldsymbol{\lambda}) = \mathbf{p}'\boldsymbol{\Theta}^2\mathbf{e}_T. \quad (21)$$

5.3 Hardware and software

The Monte Carlo simulation experiment in Section 6 and the empirical application in Section 7 focus on the special cases of MAFE (20) and MSFE (21). In these cases, the stochastic optimization problem reduce to a Linear Programming or Convex Quadratic Programming problem with limited computational complexity. The computational burden is not trivial, but remains manageable with standard computer hardware and software.

The linear and quadratic optimization problems are modeled and solved with IBM ILOG CPLEX Optimization Studio 12.8.0.0 on a computer with 2 x Intel(R) Xeon(R) CPU E5-2695 2.10GHz processors and 512GB memory. Particularly for the quadratic problems, the CPLEX’s ‘Numerical Emphasis’ option is activated to guarantee convergence to optimal solutions.

The most demanding problem in this study is the Quadratic Programming problem for minimizing MSFE subject to SCLSD constraints with a sample size of $T=1000$ in the simulation experiment. In this case, the quartile breakpoints for the solution time are 827.15s (25th percentile), 870.69s (median) and 906.39s (75th percentile). The corresponding problem is solved repeatedly for a moving estimation window of size $T = 250$ in the empirical application. In this case, the breakpoints are 2.89s (25th percentile), 3.33s (median) and 4.07s (75th percentile).

6 Simulation Experiment

The numerical example of Section 3 is extended to analyze the effect of sampling error on the optimal forecast combination and the effect of the dominance constraints on the goodness of the solution, using Monte Carlo simulation.

Again, X and Y_1 are independent standard uniform random variables. The second forecast, Y_2 , is the mean of $n = 1, 2$ additional independent standard uniform random variables. The EWA ($\lambda_1^* = \frac{1}{2}$ and $\lambda_2^* = \frac{1}{2}$) is the optimal combination for $n = 1$, while the optimal mixing weights are $\lambda_1^* = \frac{1}{3}$ and $\lambda_2^* = \frac{2}{3}$ for $n = 2$.

A total of 1000 independent random samples of $T = 10, 30, 100, 300, 1000$ paired observations of X, Y_1 and Y_2 are drawn. For every random sample, the latent CDF \mathcal{F} is simply estimated using the ECDF \mathcal{F}_T , since the paired observations are serially independent and identically distributed.

For every sample, the unconstrained empirical problem $\max_{\lambda \in \Lambda} G_{\mathcal{F}_T, \lambda}$ and its constrained counterpart (7) are solved. The constrained problem is solved using Convex Optimization problem formulation (18). The objective is minimization of the sample MSFE, using the convex quadratic goal function (20). Very similar results are obtained for minimizing of MAFE (not reported here). The slack function for the dominance constraints is set at $c_T = 10^{-3} \log(T) T^{-0.5}$.

Given the optimal solution for the mixing weights, the population MSFE can be computed as $\left((\lambda_{T,1}^*)^2 + \frac{1}{n} (\lambda_{T,2}^*)^2 + 1 \right) \frac{1}{12}$, where $n = 1, 2$ is the number of standard uniform random variables used to construct Y_2 .

Table I summarizes the simulation results. Panel A shows the case where both forecasts are equivalent ($n = 1$), the EWA ($\lambda_1^* = \frac{1}{2}$ and $\lambda_2^* = \frac{1}{2}$) is the optimal combination and the optimal MSFE is $\frac{1}{8} = 0.125$. Shown are the 10th, 25th, 50th, 75th and 90th percentile break-points (P10, ..., P90) of the weight for Y_1 ($\lambda_{T,1}^*$) and the MSFE $\left((\lambda_{T,1}^*)^2 + (\lambda_{T,2}^*)^2 + 1 \right) \frac{1}{12}$.

The unconstrained optimal weights are quite sensitive to sampling variation, witness, for example, the frequency with which $\lambda_{T,1}^* \leq \frac{1}{3}$ or $\lambda_{T,1}^* \geq \frac{2}{3}$ occurs in random samples of size $T = 30$. As a result, the MSFE frequently exceeds the minimum value of 0.125 by a significant margin in small samples. By contrast, the optimal solution with SD constraints quickly converges to the latent optimum ($\lambda_1^* = \frac{1}{2}$; $\lambda_2^* = \frac{1}{2}$; MSFE=0.125).

Panel B shows the case where Y_2 is more accurate than Y_1 ($n = 2$), the optimal combination is $\lambda_1^* = \frac{1}{3}$ and $\lambda_2^* = \frac{2}{3}$, and the optimal MSFE is $\frac{1}{9} \approx 0.111$. Again, the dominance constraints materially reduce the sensitivity to sampling variation. Naturally, the constrained optimal solution is biased towards to the non-optimal EWA benchmark in small samples. However, the effect of the bias is limited due to the relative goodness of the EWA benchmark (compared with individual forecasts) and, in addition, the bias vanishes as the sample size increases.

The simulation results confirm the statistical consistency of the mixing weights which was established in Theorem 4.4.b. The results also illustrate how the dominance constraints

help to avoid extreme solutions and improve the robustness and accuracy of the solution in finite samples.

[Insert Table I about here.]

7 Application

7.1 VIX forecasting

The proposed methodology is applied to the forecasting of daily log returns to the VIX. Using daily data ensures a large number of out-of-sample forecasts, which is favorable for the power of statistical tests aimed at distinguishing between the predictive ability of competing forecast combination methods. Furthermore, using daily data ensures frequent and recent signals for active trading.

A data set from January 02, 1990, through November 16, 2017 (7,024 trading days), is taken from the FRED database of the St. Louis Federal Reserve Bank. The first 750 days are used for creating initial forecasts and the data set used for optimization therefore starts on December 18, 1992. A separate analysis is made for the years 2007 and 2008 to account for the possible effects of the Global Financial Crisis (GFC).

Seven base forecasts are considered based on stylized facts which were documented already by Harvey and Whaley (1992) before the start of the sample period: the mean-reversal for short horizons and the decreasing pattern of market volatility as the number of remaining days in the week falls (higher VIX on Mondays and lower VIX on Friday).

Six forecasts are based on simple linear univariate regressions based on a single regressor and a 500-day moving estimation window. The six regressors are (1) one-day lagged return,

(2) moving average (MA) of the past two days, (3) MA of the past 5 days, (4) MA of the past 10 days, (5) MA of the past 20 days and (6) MA of the past 60 days. The seventh forecast is the MA in the past 500 days on the same weekday (MAW).

Since each regressor contains complementary information, combinations of the seven base forecasts can be expected to perform better than individual base forecasts in this application.

For estimating the optimal weights, the analysis uses a rolling estimation window with a length of $T = 250$ days. In each estimation window, the optimal forecast combination is constructed based on the seven base forecasts.

The objective is to minimize either MAFE or MSFE. The objective function is minimized both with and without the SD constraints (18). The slack function for the dominance constraints is set equal to $c_T = 10^{-4} \log(T)T^{-0.5}$.

The benchmark for the SD constraints is the Equal Weighted Average (EWA) of the seven forecasts. Optimality conditions for the EWA (under the population distribution) are discussed in Capistran and Timmermann (2009), Hsiao and Wan (2014) and Chan and Pauwels (2018). A possible alternative benchmark is $Y = 0$, which is motivated by the Random Walk Hypothesis. This alternative benchmark is however clearly inferior to the EWA in this application.

7.2 Forecast accuracy measurement

The forecast accuracy of the forecast combination is evaluated out of sample based on the forecast error for the first trading day after the end of the estimation window ('Day $T + 1$ ').

The first out-of-sample forecasts are made for December 15, 1993, based on an estimation window which ranges from December 18, 1992, through December 14, 1993. The next forecast is for December 16, 1994, and shifts the estimation window by one day, starting on December 21, 1992, and ending on December 15, 1993; and so forth.

To measure the out-of-sample goodness, the analysis reports the intercept (c_0), slope (c_1)

and coefficient of determination (R^2) of the Mincer-Zarnowitz (MZ) regression of actual on forecasted values, together with the Diebold and Mariano (1995) statistics for the MAFE or MSFE of the optimized combination being equal to that of the EWA benchmark. Given that one-day ahead forecasts and a long time series are used and the VIX is relatively close to a random walk process, corrections for serial correlation and small-sample bias (Harvey, Leybourne and Newbold (1997)) have a minimal effect on the DM statistics. Accounting for the nested nature of the models (Clark and West (2007)) also does not change the conclusions.

Special attention is given to the predicted sign of the VIX change. The analysis reports the fraction of correct predictions for the sign of the VIX change ('Sign'). The analysis also studies the return-reward trade-off of a simple conditional volatility trading strategy which takes a long position in the volatility when an increase is predicted and a short position when a decrease is predicted. Since 2004, volatility trading can be implemented using VIX futures contracts. Prior to 2004, delta-neutral straddles on the underlying S&P 500 index would be the tradable instrument.

The strategy is very profitable on paper, with an annualized gross Sharpe ratio of around 2 (using the EWA forecast combination). However, the returns are computed here without accounting for financing costs, transactions costs, basis risk (for futures contracts), gamma risk (for straddles) and slippage due to delayed implementation of the daily strategy and hence materially overstates the net Sharpe ratio which is feasible in practice.

Since the net returns of the strategy are difficult to estimate without considering the available trading facilities, the focus here is on the incremental effects of using optimized forecast combinations, using the (annualized) Information Ratio (IR) relative to the trading returns based on the EWA forecast.

7.3 Results

Table I summarizes the out-of-sample forecast accuracy of the five competing forecast combinations: EWA, MAFE minimization, MAFE minimization with SD constraints, MSFE

minimization and MSFE minimization with SD constraints.

EWA yields an MZ R^2 of about 2.7%. The limited forecast success is not surprising given the noise from unpredictable daily news flows and changes in market conditions at a daily horizon. The R^2 is however materially higher than the values shown in the analysis of S&P100 implied volatility for 1985-1989 in Harvey and Whaley (1992, Table 4). Furthermore, the predictive ability is potentially economically relevant, witness the aforementioned Sharpe Ratio.

The slope of the MZ regression of $c_1 = 2.143$ suggests that the EWA is biased below the realized VIX return out of sample. A closer inspection reveals that the bias stems from a significant non-linearity (concavity) of the relation between realizations and forecasts for some univariate regression models (notably those models based on short-term reversal), which makes the median forecasts of those models too conservative. The EWA reduces the dispersion of the forecasts by diversification across multiple models but it does not solve the conservative median forecasts of the univariate models, which is reflected in the conservative mean forecast of the EWA.

The minimization of MAFE and MSFE improves significantly upon the statistical goodness of EWA, witness the DM statistics for MAFE and MSFE. The MZ regressions reveal that these improvements stem mostly from reducing the bias of the EWA and the MZ R^2 does not improve materially.

The minimization of MAFE leads to higher accuracy than the minimization of MSFE. The optimal combination weights reveal that the higher accuracy is achieved by placing more emphasis on day-of-the-week effects (Y_7 is the univariate regression forecasts based on the MAW variable).

In terms of economic significance, the minimization of MAFE and MSFE leads to negative IRs which amounts to deterioration of the Sharpe ratio relative to the EWA. The IR of -0.142 for MAFE minimization is a striking contrast to the statistical goodness measures. A closer inspection reveals that the optimized models are less successful at forecasting large

VIX increases than the EWA and thus provide an inferior basis for conditional volatility trading.

The forecasting accuracy of the optimized models deteriorates sharply during the GFC subsample. In this subsample, all accuracy measures point at underperformance relative to EWA: The MZ R^2 drops below 2 percent, the DM statistics turn negative and the IRs show large negative values.

The optimized combinations strongly benefit from imposing the SD constraints by every criterion. The MZ R^2 improves materially, reflecting improvements upon EWA beyond bias correction. The DM statistics further increase and indicate highly significant reductions in MAFE and MSFE. The IRs turn from negative values to positive values. The optimal combination weights show the moderating effect of the SD constraints. The results are more robust during the GFC period; in this period, the optimal combination weights resemble those of EWA.

Whereas unconstrained optimization of forecast combinations struggles to improve upon simple averaging, robust optimization based on SD constraints yields material improvements in statistical and economic goodness in this application.

[Insert Table 1 about here.]

References

- [1] Anderson, G., 1996, Nonparametric tests of stochastic dominance in income distributions, *Econometrica*, 64, 1183-1193.
- [2] Andrews, D. W. , 1999, Estimation When a Parameter is on a Boundary, *Econometrica* 67, 6, 1341-1383.
- [3] Andrews D. W. K. and D. Pollard, 1994, An Introduction to Functional Central Limit Theorems for Dependent Stochastic Processes, *International Statistical Review* 62, 119-132.

- [4] Andrews, D. W., and G. Soares, 2010, Inference for parameters defined by moment inequalities using generalized moment selection, *Econometrica* 78, 119-157.
- [5] Back, K. and D.P. Brown, 1993, Implied Probabilities in GMM Estimators, *Econometrica* 61, 971-975.
- [6] Barrett, G. and S. Donald, 2003, Consistent tests for stochastic dominance, *Econometrica* 71, 71-104.
- [7] Bates, G.E., 1955, Joint distributions of time intervals for the occurrence of successive accidents in a generalized Polya urn scheme, *Annals of Mathematical Statistics* 26, 705-720.
- [8] Bates, JM and CWJ Granger, 1969, The Combination of Forecasts, *Operations Research Quarterly* 20, 451-468.
- [9] 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.
- [10] Beer, Gerald, 1993, *Topologies on closed and closed convex sets*, Springer Science & Business Media, vol. 268.
- [11] Beer, G., Rockafellar, R. T., and Wets, R. J. B., 1992, A characterization of epi-convergence in terms of convergence of level sets. *Proceedings of the American Mathematical Society*, 116(3), 753-761.
- [12] Bertsekas, D. P., 1999, *Nonlinear programming*, Belmont: Athena scientific.
- [13] Borwein, J., and Lewis, A. S., 2010, *Convex analysis and nonlinear optimization: theory and examples*, Springer Science & Business Media.
- [14] Capistran, C., and A. Timmermann, 2009, Forecast combination with entry and exit of experts, *Journal of Business & Economic Statistics* 27, 428-440.
- [15] Chan, F. and L.L. Pauwels, 2018, Some theoretical results on forecast combinations, *International Journal of Forecasting* 34, 64-74.
- [16] Claeskens, G., J.R. Magnus, A.L. Vasnev and W. Wang, 2016, A simple theoretical explanation of the forecast combination puzzle, *International Journal of Forecasting* 32, 754-762.
- [17] Clark, T. E. and K.D. West, 2007, Approximately normal tests for equal predictive accuracy in nested models, *Journal of Econometrics* 138, 291-311.
- [18] Clemen, R.T., 1989, Combining forecasts: A review and annotated bibliography, *International Journal of Forecasting* 5, 559-583.
- [19] Cortissoz, J., 2007, On the Skorokhod representation theorem, *Proceedings of the American Mathematical Society* 135, 3995-4007.

- [20] Dalayan, A. and J. Salmon, 2012, Sharp oracle inequalities for aggregation of affine estimators, *Annals of Statistics* 40, 2327-2355.
- [21] Davidson, R. and J.-Y. Duclos, 2000, Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality, *Econometrica* 68, 1435-1464.
- [22] Dedecker, J., and Louhichi, S., 2002, Maximal inequalities and empirical central limit theorems, in *Empirical process techniques for dependent data* (pp. 137-159), Birkhäuser, Boston, MA.
- [23] Dentcheva, D. and A. Ruszczyński, 2003, Optimization with Stochastic Dominance Constraints, *SIAM Journal on Optimization* 14, 548–566.
- [24] Diebold, F. and R. Mariano, 1995, Comparing predictive accuracy, *Journal of Business & Economic Statistics* 13, 253–265.
- [25] Gneiting, T., 2011, Making and Evaluating Point Forecasts, *Journal of the American Statistical Association* 106, 746-762.
- [26] Hadar, J. and W.R. Russell, 1969, Rules for Ordering Uncertain Prospects, *American Economic Review* 59, 2–34.
- [27] Hanoch, G. and H. Levy, 1969, The Efficiency Analysis of Choices Involving Risk, *Review of Economic Studies* 36, 335–346.
- [28] Harvey, C.R. and R.E. Whaley, 1992, Market volatility prediction and the efficiency of the S&P 100 index option market, *Journal of Financial Economics* 64, 43-73.
- [29] Harvey, D., S. Leybourne and P. Newbold, 1997, Testing the equality of prediction mean square errors, *International Journal of Forecasting* 13, 281–291.
- [30] Hsiao, C. and S. K. Wan, 2014, Is there an optimal forecast combination?, *Journal of Econometrics* 178, 294–309.
- [31] Jin, S., V. Corradi and N.R. Swanson, 2017, Robust Forecast Comparison, *Econometric Theory* 33, 1306-1351.
- [32] Juditsky, A., and Nemirovski, A., 2000, Functional aggregation for nonparametric regression, *Annals of Statistics* 28, 681-712.
- [33] Kim, J., and Pollard, D., 1990, Cube root asymptotics, *The Annals of Statistics* 18, 191-219.
- [34] Kitamura, Yuichi, 1997, Empirical likelihood methods with weakly dependent processes, *Annals of Statistics* 25, 2084-2102.
- [35] Knight, K., 1999, Epi-convergence in distribution and stochastic equi-semicontinuity. Unpublished manuscript, 37, 28-72.

- [36] Kuosmanen, T., 2004, Efficient diversification according to stochastic dominance criteria, *Management Science* 50, 1390-1406.
- [37] Lavancier, F. and P. Rochet, 2016, A general procedure to combine estimators, *Computational Statistics & Data Analysis* 94, 175-192.
- [38] 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.
- [39] Linton, O.B., Th. Post and Y.-J. Whang, 2014, Testing for the Stochastic Dominance Efficiency of a Given Portfolio, *Econometrics Journal* 17, 59-74.
- [40] McCracken, M.W., 2000, Robust out-of-sample inference, *Journal of Econometrics* 99, 195-223.
- [41] 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.
- [42] Molchanov, I., 2006, *Theory of random sets*. Springer Science & Business Media.
- [43] Newbold, P. and C.W.J. Granger, 1974, Experience with Forecasting Univariate Time Series and the Combination of Forecasts, *Journal of the Royal Statistical Society A* 137, 131-165.
- [44] Patton, Andrew J., 2011, Volatility forecast comparison using imperfect volatility proxies, *Journal of Econometrics* 160, 246-256.
- [45] Poon, S-H. & C.W.J. Granger, 2003, Forecasting volatility in financial markets: a review, *Journal of Economic Literature* 41, 478-539.
- [46] Post, T., S. Karabati and S. Arvanitis, 2018, Portfolio Construction Based on Stochastic Dominance and Empirical Likelihood, *Journal of Econometrics* 206, 167-186.
- [47] Rigollet, Ph and A.B. Tsybakov, 2007, Linear and convex aggregation of density estimators, *Mathematical Methods of Statistics* 15, 260-280.
- [48] Rio, E., 2013, Inequalities and Limit Theorems for Weakly Dependent Sequences, Lecture Notes available at <https://cel.archives-ouvertes.fr/cel-00867106/document>.
- [49] Rockafellar R.T. and S. Uryasev, 2000, Optimization of Conditional Value-at-Risk, *Journal of Risk* 2, 21-41.
- [50] Rockafellar, R. T., & Wets, R. J. B., 2009, *Variational analysis* (Vol. 317). Springer Science & Business Media.
- [51] Roman, D., K. Darby-Dowman and G. Mitra, 2006, Portfolio construction based on stochastic dominance and target return distributions, *Mathematical Programming* 108, 541-569.

- [52] Rothschild, M. and J. E. Stiglitz, 1970, Increasing Risk: I. A Definition, *Journal of Economic Theory* 2, 225-243.
- [53] Qin, J. and J. Lawless, 1994, Empirical likelihood and general estimating equations, *Annals of Statistics* 22, 300–325.
- [54] Scaillet, O., and N. Topaloglou, 2010, Testing for stochastic dominance efficiency, *Journal of Business & Economic Statistics* 28, 169-180.
- [55] Schmid, F. and Trede, M., 1998, A Kolmogorov-type test for second order stochastic dominance, *Statist. Probab. Lett.* 37, 183-193.
- [56] Smith, J. and K. Wallis, 2009, A Simple Explanation of the Forecast Combination Puzzle, *Oxford Bulletin of Economics and Statistics* 71, 331–355.
- [57] West, K. D., and McCracken, M. W., 1998, Regression-based tests of predictive ability, *International Economic Review* 39, 817-840.
- [58] van der Vaart, A.W., 2000, *Asymptotic Statistics*, Cambridge University Press.

Appendix

Proof of Proposition 4.3. Consider the event $\mathcal{A}_T := \{\boldsymbol{\lambda} \notin \Lambda_{\mathcal{F}_T}^{\succ} (c_T), \forall \boldsymbol{\lambda} \notin \Lambda_{\mathcal{F}}^{\succ}\}$. Due to Lemma 4.2 and $\sup_{z \in \mathcal{Z}^-} c_T(z) \rightarrow 0$, it is found that $\mathbb{P}(\mathcal{A}_T) \rightarrow 1$, as $T \rightarrow \infty$. By construction $\mathcal{A}_T \subseteq \{\boldsymbol{\lambda}_T \in \Lambda_{\mathcal{F}}^{\succ}\}$ and the result follows since by construction $\boldsymbol{\lambda}_T \in \Lambda_{\mathcal{F}_T}^{\succ} (c_T)$. ■

Proof of Theorem 4.4. In what follows $\mathcal{D}_{\boldsymbol{\lambda}}(\mathcal{F}, z) := \mathbb{E}_{\mathcal{F}}[L_z(\mathbf{U}\boldsymbol{\tau})] - \mathbb{E}_{\mathcal{F}}[L_z(\mathbf{U}\boldsymbol{\lambda})]$ is the achieved reduction of expected loss and $\mathcal{Z}_{\boldsymbol{\lambda}}^- := \{z \in \mathcal{Z}^- : \mathcal{D}_{\boldsymbol{\lambda}}(z) = 0\}$ the ‘contact set’ for $\boldsymbol{\lambda}$. Assumption 4.1 and the Continuous Mapping Theorem imply that $\mathcal{D}_{\boldsymbol{\lambda}}(\sqrt{T}(\mathcal{F}_T - \mathcal{F}), z)$ weakly converges to $\mathcal{G}(\boldsymbol{\tau}, z) - \mathcal{G}(\boldsymbol{\lambda}, z)$ with respect to to the product topology of continuous (w.r.t. $\boldsymbol{\lambda}$) epi-convergence (w.r.t. z) on the product of the relevant spaces of usc real valued functions (see, e.g., Knight (1999)). This product space is metrizable as complete and separable (see again Knight (1999)). Hence, via Skorokhod representations which are applicable due to Cortissoz (2007, Thm 1) and by using the same notation, for any $\boldsymbol{\lambda}$ and any sequence $\boldsymbol{\lambda}_T \rightarrow \boldsymbol{\lambda}$, it follows that $\mathcal{D}_{\boldsymbol{\lambda}_T}(\sqrt{T}(\mathcal{F}_T - \mathcal{F}), z)$ converges to $\mathcal{G}(\boldsymbol{\tau}, z) - \mathcal{G}(\boldsymbol{\lambda}, z)$ almost surely, w.r.t. to the topology of epi-convergence. Notice that,

$$\mathcal{D}_{\boldsymbol{\lambda}}(\sqrt{T}\mathcal{F}_T, z) = \mathcal{D}_{\boldsymbol{\lambda}}(\sqrt{T}(\mathcal{F}_T - \mathcal{F}), z) + \sqrt{T}\mathcal{D}_{\boldsymbol{\lambda}}(\mathcal{F}, z),$$

and since $\mathcal{D}_{\boldsymbol{\lambda}}(\mathcal{F}, \cdot)$ is locally Lipschitz locally uniformly w.r.t. $\boldsymbol{\lambda}$ imply that almost surely, $\mathcal{D}_{\boldsymbol{\lambda}}(\sqrt{T}\mathcal{F}_T, z)$ converges w.r.t. to the product topology of continuous (w.r.t. $\boldsymbol{\lambda}$) hypo-convergence (w.r.t. z) to

$$\Gamma(\boldsymbol{\lambda}, z) := \begin{cases} \mathcal{G}(\boldsymbol{\tau}, z) - \mathcal{G}(\boldsymbol{\lambda}, z), & \boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}^{\succ}, z \in \mathcal{Z}_{\boldsymbol{\lambda}}^- \\ +\infty, & \boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}^{\succ}, z \notin \mathcal{Z}_{\boldsymbol{\lambda}}^- \\ -\infty, & \boldsymbol{\lambda} \notin \Lambda_{\mathcal{F}}^{\succ} \end{cases}$$

Then, Molchanov (2006, Prop 3.2, Ch 5, p. 337) and the relationship between epi-convergence and continuous convergence (see, e.g., Rockafellar and Wets (2009)) imply that $\inf_{z \in \mathcal{Z}^-} \mathcal{D}_\lambda \left(\sqrt{T} \mathcal{F}_T, z \right)$ hypoconverges almost surely to $\inf_{z \in \mathcal{Z}^-} \Gamma(\boldsymbol{\lambda}, z)$. Notice that

$$\Lambda_{\mathcal{F}_T}^{\succeq}(c_T) \subseteq \left\{ \boldsymbol{\lambda} \in \Lambda : \inf_{z \in \mathcal{Z}^-} \mathcal{D}_\lambda \left(\sqrt{T} \mathcal{F}_T, z \right) \geq -\sqrt{T} \sup_{z \in \mathcal{Z}^-} c_T \right\}$$

and analogously

$$\Lambda_{\mathcal{F}_T}^{\succ} = \left\{ \boldsymbol{\lambda} \in \Lambda : \inf_{z \in \mathcal{Z}^-} \Gamma(\boldsymbol{\lambda}, z) > -\infty \right\}, \text{ almost surely.}$$

Using this and (the dual part of) Beer, Rockafellar and Wets (1992, Thm 3.1), the result in part (a) follows, because, for any $\boldsymbol{\lambda} \in \Lambda_{\mathcal{F}}^{\succeq}$ for which $\mathcal{Z}_\lambda^- \neq \emptyset$, due to the Portmanteau Lemma and the properties of c_T ,

$$\mathbb{P} \left(\inf_{z \in \mathcal{Z}^-} \mathcal{D}_\lambda \left(\sqrt{T} \mathcal{F}_T, z \right) \geq -\sqrt{T} \sup_{z \in \mathcal{Z}^-} c_T(z) \right) \rightarrow \mathbb{P} \left(\inf_{z \in \mathcal{Z}_\lambda^-} (\mathcal{G}(\boldsymbol{\tau}, z) - \mathcal{G}(\boldsymbol{\lambda}, z)) \geq -\infty \right)$$

which equals to one, while if $\boldsymbol{\lambda}_T \rightsquigarrow \boldsymbol{\lambda} \notin \Lambda_{\mathcal{F}}^{\succeq}$ (or the latter is a cluster point), since $\Lambda_{\mathcal{F}}^{\succ}$ is open and by construction $\inf_{z \in \mathcal{Z}^-} \mathcal{D}_{\boldsymbol{\lambda}_T} \left(\sqrt{T} \mathcal{F}, z \right)$ almost surely diverges to $-\infty$ faster than $-\sqrt{T} \sup_{z \in \mathcal{Z}^-} c_T(z)$, it follows that any such $\boldsymbol{\lambda}$ cannot lie inside the limiting set. For part (b), notice that, since Λ is compact and separable, the Painleve-Kuratowski convergence on the hyperspace of its closed non-empty subsets is metrizable (see Molchanov (2006, Appendix B)) and the relevant metric space is complete and separable. Lemma 4.2, implies as previously the epi-convergence in probability of $\sum_{t=1}^T p_t L(X_t - \mathbf{Y}_t \boldsymbol{\lambda})$ to $\mathbb{E}(L(X_t - \mathbf{Y}_t \boldsymbol{\lambda}))$ as functions defined on Λ . Cortissoz (2007, Thm 1) then implies joint Skorokhod representations for both the sets involved in the convergence in part (a) and the functions involved in the previous one. This and the compactness of Λ enables again the use of Molchanov (2006, Prop 3.2, Ch. 5, p. 337) and then part (b) follows from Lemma 4.2, and part (c) follows from the fact that $\mathbb{E}(L(X_t - \mathbf{Y}_t \boldsymbol{\lambda}))$ has a unique minimizer together with Rockafellar and Wets (2009, Thm 7.31). ■

Proof of Theorem 4.6. It follows from $\boldsymbol{\lambda}_T \rightsquigarrow \boldsymbol{\lambda}_{\mathcal{F}}^*$ that $\boldsymbol{\lambda}_T \in \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}$ with probability tending to one (w.h.p.). Due to Lemma 4.2, the Continuous Mapping Theorem, the strict convexity of $\mathbb{E}_{\mathcal{F}}[L(\mathbf{U}\boldsymbol{\lambda})]$, and the Portmanteau Theorem, it follows that, for any $\varepsilon > 0$,

$$\lim_{T \rightarrow \infty} \mathbb{P} \left(\min_{\boldsymbol{\lambda} \in \Lambda \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta)} \mathbb{E}_{\mathcal{F}_T}[L(\mathbf{U}\boldsymbol{\lambda})] < \min_{\boldsymbol{\lambda} \in \Lambda_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\varepsilon)} \mathbb{E}_{\mathcal{F}_T}[L(\mathbf{U}\boldsymbol{\lambda})] \right) = 1,$$

where $\Lambda_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\varepsilon) = \left\{ \boldsymbol{\lambda} \in \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta) \cap \Lambda_{\mathcal{F}}^{\succ} : |\boldsymbol{\lambda} - \boldsymbol{\lambda}_{\mathcal{F}}^*| > \varepsilon \right\}$ and $\Lambda_{\mathcal{F}}^{\succ}$ is the complement of $\Lambda_{\mathcal{F}}^{\succeq}$ in Λ . Hence, with w.h.p. $\boldsymbol{\lambda}_T$ belongs to $\Lambda_{\mathcal{F}}^{\succeq}$ and asymptotically solves

$$\begin{aligned} & \min \mathbb{E}_{\mathcal{F}_T}[L(\mathbf{U}\boldsymbol{\lambda})] \\ \text{s.t. } & \mathbb{E}_{\mathcal{F}}[L_z(\mathbf{U}\boldsymbol{\lambda})] \leq \mathbb{E}_{\mathcal{F}}[L_z(\mathbf{U}\boldsymbol{\tau})] \quad \forall z \in \mathcal{Z}^-. \\ & \boldsymbol{\lambda} \in \Lambda \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta) \end{aligned} \tag{22}$$

This results and Assumption 4.5 imply that $\mathbf{u}_T := \sqrt{T}(\boldsymbol{\lambda}_T - \boldsymbol{\lambda}_{\mathcal{F}}^*)$ is the minimizer of

$$T \left(\mathbb{E}_{\mathcal{F}_T} \left[L \left(\mathbf{U} \left(\boldsymbol{\lambda}_{\mathcal{F}}^* + \frac{1}{\sqrt{T}} \mathbf{u} \right) \right) \right] - \mathbb{E}_{\mathcal{F}_T} [L(\mathbf{U}(\boldsymbol{\lambda}_{\mathcal{F}}^*))] \right)$$

over $\sqrt{T}(\Lambda_{\mathcal{F}}^{\succeq} \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta) - \boldsymbol{\lambda}_{\mathcal{F}}^*)$.

Assume first that $\mathbb{E}_{\mathcal{F}}[s_0(\boldsymbol{\lambda}_{\mathcal{F}}^*)] = \mathbf{0}_M$. This assumption together with Assumptions 4.5.(i)-(ii) imply that, for $\mu_{\mathcal{F}}^*$ necessarily converging to $\boldsymbol{\lambda}_{\mathcal{F}}^*$ w.h.p.,

$$\sqrt{T} \mathbb{E}_{\mathcal{F}_T} [s_t(\boldsymbol{\lambda}_{\mathcal{F}}^*)]' \mathbf{u}_T + \frac{1}{2} \mathbf{u}'_T \mathbb{E}_{\mathcal{F}_T} [H_t(\mu_{\mathcal{F}}^*)] \mathbf{u}_T \leq O_p(1),$$

hence

$$\sqrt{T} \mathbb{E}_{\mathcal{F}_T} [s_t(\boldsymbol{\lambda}_{\mathcal{F}}^*)]' \mathbf{u}_T + \frac{1}{2} \mathbf{u}'_T (H_{\boldsymbol{\lambda}_{\mathcal{F}}^*} + o_p(1)) \mathbf{u}_T \leq O_p(1),$$

which then, due to Assumption 4.5.(ii), implies that

$$(\|\mathbf{u}_T\|^2 + 2\|\mathbf{u}_T\|)O_p(1)(1 + o_p(1)) + O_p(1) \leq O_p(1),$$

and thereby $\|\mathbf{u}_T\|(1 + o_p(1)) \leq O_p(1)$, which establishes uniform tightness for \mathbf{u}_T . Due to Beer (1993, Exercise 10 in Exercise Set 5.2), $\sqrt{T}(\Lambda_{\mathcal{F}}^{\succeq} \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta) - \boldsymbol{\lambda}_{\mathcal{F}}^*)$ converges in the Painleve-Kuratowski sense to the non-empty, closed and convex $\Lambda_{\infty}^{\succeq}$, while, due to Assumption 4.5.(ii), the sup distance between $\sqrt{T} \mathbb{E}_{\mathcal{F}_T} [s_t(\boldsymbol{\lambda}_{\mathcal{F}}^*)]' \mathbf{u} + \frac{1}{2} \mathbf{u}' \mathbb{E}_{\mathcal{F}_T} [H_t(\boldsymbol{\lambda}_{\mathcal{F}}^*)] \mathbf{u}$ and $\sqrt{T} \mathbb{E}_{\mathcal{F}_T} [s_t(\boldsymbol{\lambda}_{\mathcal{F}}^*)]' \mathbf{u} + \frac{1}{2} \mathbf{u}' \mathbb{E}_{\mathcal{F}_T} [H_t(\mu_{\mathcal{F}}^*)] \mathbf{u}$ is $o_p(1)$, locally uniformly on $\sqrt{T}(\Lambda_{\mathcal{F}}^{\succeq} \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta) - \boldsymbol{\lambda}_{\mathcal{F}}^*)$. Furthermore, for general $\mathbf{u} \in \mathbb{R}^M$, the function $\sqrt{T} \mathbb{E}_{\mathcal{F}_T} [s_t(\boldsymbol{\lambda}_{\mathcal{F}}^*)]' \mathbf{u} + \frac{1}{2} \mathbf{u}' \mathbb{E}_{\mathcal{F}_T} [H_t(\boldsymbol{\lambda}_{\mathcal{F}}^*)] \mathbf{u}$ has the same minimizers as $\left\| \mathbb{E}_{\mathcal{F}_T}^{\frac{1}{2}} [H_t(\boldsymbol{\lambda}_{\mathcal{F}}^*)] \mathbf{u} + \mathbb{E}_{\mathcal{F}_T}^{-\frac{1}{2}} [H_t(\boldsymbol{\lambda}_{\mathcal{F}}^*)] \sqrt{T} \mathbb{E}_{\mathcal{F}_T} [s_t(\boldsymbol{\lambda}_{\mathcal{F}}^*)] \right\|$ and the latter weakly converges to $\left\| \mathcal{H}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}^{\frac{1}{2}} \mathbf{u} + \mathcal{H}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}^{-\frac{1}{2}} S_{\boldsymbol{\lambda}_{\mathcal{F}}^*} \right\|$. Hence, van der Vaart (2000, Cor 5.58) applies, which establishes the sought result.

Assume now that $\mathbb{E}_{\mathcal{F}}[s_0(\boldsymbol{\lambda}_{\mathcal{F}}^*)] \neq \mathbf{0}_M$. Assumption 4.5.(iii), convexity of $\Lambda_{\mathcal{F}}^{\succeq} \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta)$, convexity of expected loss and Borwein and Lewis (2010, Prop 6.3.9) imply that $\boldsymbol{\lambda}_{\mathcal{F}}^*$ is the unique solution of (4) iff $-\mathbb{E}_{\mathcal{F}}[s_0(\boldsymbol{\lambda}_{\mathcal{F}}^*)]$ belongs to the normal cone $\mathcal{N}(\boldsymbol{\lambda}_{\mathcal{F}}^*, \Lambda_{\mathcal{F}}^{\succeq} \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta))$ and hence $\mathcal{N}(\mathbf{0}_M, \sqrt{T}(\Lambda_{\mathcal{F}}^{\succeq} \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta) - \boldsymbol{\lambda}_{\mathcal{F}}^*))$ and $\mathcal{N}(\mathbf{0}_M, \Lambda_{\infty}^{\succeq})$. Furthermore, $\mathbf{v}_T := \sqrt{T} \mathbf{u}_T$ minimizes the function $\mathbb{E}_{\mathcal{F}_T} [L(\mathbf{U}(\boldsymbol{\lambda}_{\mathcal{F}}^* + \mathbf{v}))] - \mathbb{E}_{\mathcal{F}_T} [L(\mathbf{U}(\boldsymbol{\lambda}_{\mathcal{F}}^*))]$ over $\Lambda_{\mathcal{F}}^{\succeq} \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta) - \boldsymbol{\lambda}_{\mathcal{F}}^*$, which, due to Assumption 4.5.(i), equals $\mathbb{E}_{\mathcal{F}_T} [s_t(\boldsymbol{\lambda}_{\mathcal{F}}^*)]^T \mathbf{v} + \frac{1}{2} \mathbf{v}' \mathbb{E}_{\mathcal{F}_T} [H_t(\mu_{\mathcal{F}}^*)] \mathbf{v}$, where $\mu_{\mathcal{F}}^*$ is as above. Due to Assumption 4.5.(iii), the latter weakly converges locally uniformly to the strictly convex $\frac{1}{2} \left(\mathbf{v} + H_{\boldsymbol{\lambda}_{\mathcal{F}}^*}^{-1} \mathbb{E}_{\mathcal{F}} [s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*)] \right)' H_{\boldsymbol{\lambda}_{\mathcal{F}}^*} \left(\mathbf{v} + H_{\boldsymbol{\lambda}_{\mathcal{F}}^*}^{-1} \mathbb{E}_{\mathcal{F}} [s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*)] \right)$. This function is uniquely minimized at $-H_{\boldsymbol{\lambda}_{\mathcal{F}}^*}^{-1} \mathbb{E}_{\mathcal{F}} [s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*)]$, which is then projected to $\Lambda_{\mathcal{F}}^{\succeq} \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta) - \boldsymbol{\lambda}_{\mathcal{F}}^*$ to obtain its constrained minimizer there. Theorem 4.4.(c) implies that the latter is zero, which implies, via Borwein and Lewis (2010, Prop 6.3.9), that $-H_{\boldsymbol{\lambda}_{\mathcal{F}}^*}^{-1} \mathbb{E}_{\mathcal{F}} [s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*)]$ must also lie inside the aforementioned normal cones. Similarly, \mathbf{u}_T is w.h.p. the minimizer of $\kappa'_T \mathbf{u} + \sqrt{T} \mathbb{E}_{\mathcal{F}} [s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*)]' \mathbf{u} + \frac{1}{2} \mathbf{u}' \mathbb{E}_{\mathcal{F}_T} [H_t(\mu_{\mathcal{F}}^*)] \mathbf{u}$ over $\sqrt{T}(\Lambda_{\mathcal{F}}^{\succeq} \cap \bar{B}_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\eta) - \boldsymbol{\lambda}_{\mathcal{F}}^*)$, where $\kappa_T := \sqrt{T}(\mathbb{E}_{\mathcal{F}_T} [s_t(\boldsymbol{\lambda}_{\mathcal{F}}^*)] - \mathbb{E}_{\mathcal{F}} [s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*)])$. Due to Assumption 4.5.(ii), the function is convex w.h.p. Its unconstrained minimizer is w.h.p. equal to $-\mathbb{E}_{\mathcal{F}_T}^{-1} [H_t(\mu_{\mathcal{F}}^*)] \kappa_T - \mathbb{E}_{\mathcal{F}_T}^{-1} [H_t(\mu_{\mathcal{F}}^*)] \sqrt{T} \mathbb{E}_{\mathcal{F}} [s^*(\boldsymbol{\lambda}_{\mathcal{F}}^*)]$,

which is then projected to $\sqrt{T} (\Lambda_{\mathcal{F}}^{\succ} \cap \bar{B}_{\lambda_{\mathcal{F}}^*}(\eta) - \lambda_{\mathcal{F}}^*)$. Hence, w.h.p. \mathbf{u}_T satisfies the equation $P \left(\mathbf{u} + \mathbb{E}_{\mathcal{F}_T}^{-1} [H_t(\mu_{\mathcal{F}}^*)] \kappa_T + \mathbb{E}_{\mathcal{F}_T}^{-1} [H_t(\mu_{\mathcal{F}}^*)] \sqrt{T} \mathbb{E}_{\mathcal{F}} [s^*(\lambda_{\mathcal{F}}^*)], \sqrt{T} (\Lambda_{\mathcal{F}}^{\succ} \cap \bar{B}_{\lambda_{\mathcal{F}}^*}(\eta) - \lambda_{\mathcal{F}}^*) \right) = \mathbf{u}_T$ and thereby, using the Projection Theorem (Bertsekas (1999, Prop 2.1.3)), \mathbf{u}_T equals $P \left(-\mathbb{E}_{\mathcal{F}_T}^{-1} [H_t(\mu_{\mathcal{F}}^*)] \kappa_T - \mathbb{E}_{\mathcal{F}_T}^{-1} [H_t(\mu_{\mathcal{F}}^*)] \sqrt{T} \mathbb{E}_{\mathcal{F}} [s^*(\lambda_{\mathcal{F}}^*)], \sqrt{T} (\Lambda_{\mathcal{F}}^{\succ} \cap \bar{B}_{\lambda_{\mathcal{F}}^*}(\eta) - \lambda_{\mathcal{F}}^*) \right)$ w.h.p. where, for any $u \in \mathbb{R}^M$ and C a closed convex subset, $P(u, C) := \arg \min_{z \in C} \|u - z\|$. Due to the local equality to a cone of $\Lambda_{\mathcal{F}}^{\succ} \cap \bar{B}_{\lambda_{\mathcal{F}}^*}(\eta) - \lambda_{\mathcal{F}}^*$ and Bertsekas (1999, Prop 2.1.3.(b)), \mathbf{u}_T also equals, w.h.p., $P \left(-\frac{1}{\sqrt{T}} \mathbb{E}_{\mathcal{F}_T}^{-1} [H_t(\mu_{\mathcal{F}}^*)] \kappa_T - \mathbb{E}_{\mathcal{F}_T}^{-1} [H_t(\mu_{\mathcal{F}}^*)] \mathbb{E}_{\mathcal{F}} [s^*(\lambda_{\mathcal{F}}^*)], \sqrt{T} (\Lambda_{\mathcal{F}}^{\succ} \cap \bar{B}_{\lambda_{\mathcal{F}}^*}(\eta) - \lambda_{\mathcal{F}}^*) \right)$. Given the tightness condition in Assumption 4.5.(iii), consider an arbitrary convergent subsequence of \mathbf{u}_T weakly converging, say, to \mathbf{u} . Due van der Vaart (2000, Lemma 7.13) and Assumption 4.5.(ii), the projection above converges to $P \left(-H_{\lambda_{\mathcal{F}}^*}^{-1} \mathbb{E}_{\mathcal{F}} [s^*(\lambda_{\mathcal{F}}^*)], \Lambda_{\infty}^{\succ} \right)$. The normal cone and the local equality to a cone properties then imply $\mathbf{u} = \mathbf{0}_M$. ■

Proof of Lemma 4.7. Assumption 4.1'.(iii) and Rio (2013, Cor 4.1) imply that as $T \rightarrow \infty$,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T [L_z(K(\mathbf{Z}_t, \beta) \boldsymbol{\lambda}) - G(z, \boldsymbol{\lambda}, \beta)] \xrightarrow{\text{fidi}} \mathcal{G}^*(z, \boldsymbol{\lambda}, \beta), \text{ in } \ell^\infty(\mathcal{Z}^- \times \boldsymbol{\Lambda} \times \bar{B}_{\beta_0}(\eta)),$$

where $\xrightarrow{\text{fidi}}$ denotes fidi convergence, while $\mathcal{G}^*(z, \boldsymbol{\lambda}, \beta)$ is a zero-mean Gaussian process with uniformly continuous sample paths in $\ell^\infty(\mathcal{Z}^- \times \boldsymbol{\Lambda} \times \bar{B}_{\beta_0}(\eta))$, and covariance kernel

$$K_{\mathcal{G}^*}((z_1, \boldsymbol{\lambda}_1, \beta_1), (z_2, \boldsymbol{\lambda}_2, \beta_2)) = \sum_{i=0}^{\infty} \kappa_i \text{Cov}_{\mathcal{F}}(L_{z_1}(K(\mathbf{Z}_0, \beta_1) \boldsymbol{\lambda}_1), L_{z_2}(K(\mathbf{Z}_i, \beta_2) \boldsymbol{\lambda}_2)),$$

for $\kappa_i = \begin{cases} 1, & i = 0 \\ 2, & i > 0 \end{cases}$. Assumption 4.1'.(iv), the fact that the composition of Lipschitz continuous functions is Lipschitz continuous with multiplicative Lipschitz coefficients and Andrews and Pollard (1994, Thm 2.2)-see also the discussion immediately before and after the result-imply asymptotic tightness for $\frac{1}{\sqrt{T}} \sum_{t=1}^T [L_z(K(\mathbf{Z}_t, \beta) \boldsymbol{\lambda}) - G(z, \boldsymbol{\lambda}, \beta)]$ and thereby the applicability Van Der Vaart (2000, Thm 18.14) which shows that actually

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T [L_z(K(\mathbf{Z}_t, \beta) \boldsymbol{\lambda}) - G(z, \boldsymbol{\lambda}, \beta)] \rightsquigarrow \mathcal{G}^*(z, \boldsymbol{\lambda}, \beta), \text{ in } \ell^\infty(\mathcal{Z}^- \times \boldsymbol{\Lambda} \times \bar{B}_{\beta_0}(\eta)). \quad (23)$$

Assumption 4.1'.(i)-(ii) and (23) then imply that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \left[L_z(\hat{U}_t \boldsymbol{\lambda}) - (\mathbb{E}_{\mathcal{F}} [L_z(\mathbf{U} \boldsymbol{\lambda})])|_{\mathbf{U}=\hat{U}_t} \right] \rightsquigarrow \mathcal{G}^*(z, \boldsymbol{\lambda}, \beta_0), \text{ in } \ell^\infty(\mathcal{Z}^- \times \boldsymbol{\Lambda}), \quad (24)$$

and notice that

$$K_{\mathcal{G}^*}((z_1, \boldsymbol{\lambda}_1, \beta_0), (z_2, \boldsymbol{\lambda}_2, \beta_0)) = \sum_{i=0}^{\infty} \kappa_i \text{Cov}_{\mathcal{F}}(L_{z_1}(\mathbf{U}_0 \boldsymbol{\lambda}_1), L_{z_2}(\mathbf{U}_i \boldsymbol{\lambda}_2)).$$

Now, Assumption 4.1'.(ii),(iv) and the Mean Value Theorem imply that as $T \rightarrow \infty$, almost surely,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T [(\mathbb{E}_{\mathcal{F}} [L_z(\mathbf{U} \boldsymbol{\lambda})])|_{\mathbf{U}=\hat{U}_t} - G(z, \boldsymbol{\lambda}, \beta)] = \frac{1}{\sqrt{R_T T}} \sum_{t=1}^T [D_{\beta} G(z, \boldsymbol{\lambda}, \beta_t^*)] \sqrt{R_T} (\beta_t - \beta_0),$$

with $\beta_t := [\beta_{t_1} \dots \beta_{t_M}]'$, and β_t^* a random point on the ray that connects β_t and β_0 inside $\bar{B}_{\beta_0}(\eta)$. Due to Assumption 4.1'.5 it follows that $\sup_{\mathcal{Z}^- \times \Lambda \times \bar{B}_{\beta_0}(\eta)} \|D_\beta G(z, \boldsymbol{\lambda}, \beta)\| < +\infty$, and then from Assumption 4.1'.(i),(ii),(iii),(iv) it follows that via Van Der Vaart (2000, Thm 18.14), and similar considerations to the proofs of West and McCracken (1998, Lemmata 4.1-2), that jointly with (23),

$$\frac{1}{\sqrt{R_T T}} \sum_{t=1}^T [D_\beta G(z, \boldsymbol{\lambda}, \beta)] \sqrt{R_T} (\beta_t - \beta_0) \rightsquigarrow \mathcal{G}_*(z, \boldsymbol{\lambda}, \beta), \text{ in } \ell^\infty(\mathcal{Z}^- \times \Lambda \times \bar{B}_{\beta_0}(\eta)) \quad (25)$$

where now $\mathcal{G}_*(z, \boldsymbol{\lambda}, \beta)$ is a zero-mean Gaussian process with uniformly continuous sample paths in $\ell^\infty(\mathcal{Z}^- \times \Lambda \times \bar{B}_{\beta_0}(\eta))$, and covariance kernel

$$K_{\mathcal{G}_*}((z_1, \boldsymbol{\lambda}_1, \beta_1), (z_2, \boldsymbol{\lambda}_2, \beta_2)) = \varrho_* D_\beta G(z_1, \boldsymbol{\lambda}_1, \beta_1) \mathbf{H} V_h \mathbf{H}' D_\beta G(z_2, \boldsymbol{\lambda}_2, \beta_2) .$$

The definition of β_t^* and (25) then imply that jointly with (24),

$$\frac{1}{\sqrt{R_T} \sqrt{T}} \sum_{t=1}^T [D_\beta G(z, \boldsymbol{\lambda}, \beta_t^*)] \sqrt{R_T} (\beta_t - \beta_0) \rightsquigarrow \mathcal{G}_*(z, \boldsymbol{\lambda}, \beta_0), \text{ in } \ell^\infty(\mathcal{Z}^- \times \Lambda) . \quad (26)$$

The result then follows via the Continuous Mapping Theorem for $\mathcal{G}(z, \boldsymbol{\lambda}) = \mathcal{G}^*(z, \boldsymbol{\lambda}, \beta_0) + \mathcal{G}_*(z, \boldsymbol{\lambda}, \beta_0)$ by also noticing that considerations similar to those in the proofs of West and McCracken (1998, Lemmata 4.1-2) imply

$$\text{Cov}(\mathcal{G}^*(z_1, \boldsymbol{\lambda}_1, \beta_0), \mathcal{G}_*(z_2, \boldsymbol{\lambda}_2, \beta_0)) = \varrho \sum_{i=0}^{\infty} \kappa_i \text{Cov}_{\mathcal{F}}(L_{z_1}(\mathbf{U}_0 \boldsymbol{\lambda}_1), D_\beta G(z_2, \boldsymbol{\lambda}_2, \beta_0) \mathbf{H} \mathbf{h}_i) .$$

■

Proof of Lemma 4.8. In the present proof $\ell^\infty(A, \mathbb{R}^M)$ denotes the space of $(\mathbb{R}^M, \|\cdot\|)$ -valued bounded functions on a set A equipped with the sup norm. The first part closely resembles the proof of Lemma 4.2. In this respect, Assumption 4.5'.(i)-(ii) and Rio (2013, Cor 4.1) imply that as $T \rightarrow \infty$,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \left[\frac{dL}{dx} (K(\mathbf{Z}_0, \beta) \boldsymbol{\lambda}_{\mathcal{F}}^*) K(\mathbf{Z}_0, \beta) - Q_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\beta) \right] \xrightarrow{\text{fidi}} \mathcal{G}_s(\beta), \text{ in } \ell^\infty(\bar{B}_{\beta_0}(\eta), \mathbb{R}^M),$$

where $\xrightarrow{\text{fidi}}$ denotes fidi convergence, while $\mathcal{G}^*(z, \boldsymbol{\lambda}, \beta)$ is a zero-mean Gaussian process with uniformly continuous sample paths in $\ell^\infty(\bar{B}_{\beta_0}(\eta), \mathbb{R}^M)$, and covariance kernel

$$K_{\mathcal{G}_s}(\beta_1, \beta_2) = \sum_{i=0}^{\infty} \kappa_i \text{Cov}_{\mathcal{F}} \left(\frac{dL}{dx} (K(\mathbf{Z}_0, \beta_1) \boldsymbol{\lambda}_{\mathcal{F}}^*) K(\mathbf{Z}_0, \beta_1), \frac{dL}{dx} (K(\mathbf{Z}_i, \beta_2) \boldsymbol{\lambda}_{\mathcal{F}}^*) K(\mathbf{Z}_i, \beta_2) \right),$$

for κ_i as above. Assumption 4.5'.(iii), and Andrews and Pollard (1994, Thm 2.2) imply tightness and thereby the applicability of Van Der Vaart (2000, Thm 18.14) and the Cramer-Wold device, which shows that actually

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \left[\frac{dL}{dx} (K(\mathbf{Z}_0, \beta) \boldsymbol{\lambda}_{\mathcal{F}}^*) K(\mathbf{Z}_0, \beta) - Q_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\beta) \right] \rightsquigarrow \mathcal{G}_s(\beta), \text{ in } \ell^\infty(\bar{B}_{\beta_0}(\eta), \mathbb{R}^M) . \quad (27)$$

Assumption 4.5'.(ii), the fact that $K(\mathbf{Z}_t, \beta_t) = \hat{\mathbf{U}}_t$, $K(\mathbf{Z}_t, \beta_0) = \mathbf{U}_t$, and (27) then imply that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \left[\frac{dL}{dx} \left(\hat{\mathbf{U}}_t \boldsymbol{\lambda}_{\mathcal{F}}^* \right) \hat{\mathbf{U}}_t - Q_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\beta) |_{\beta=\beta_t} \right] \rightsquigarrow \mathcal{G}_S(\beta_0), \text{ in } \ell^\infty(\mathcal{Z}^- \times \boldsymbol{\Lambda}, \mathbb{R}^M), \quad (28)$$

and notice that

$$K_{\mathcal{G}_*}(\beta_0, \beta_0) = \sum_{i=0}^{\infty} \kappa_i \text{Cov}_{\mathcal{F}} \left(\frac{dL}{dx}(\mathbf{U}_0 \boldsymbol{\lambda}_{\mathcal{F}}^*) \mathbf{U}_0, \frac{dL}{dx}(\mathbf{U}_i \boldsymbol{\lambda}_{\mathcal{F}}^*) \mathbf{U}_i \right).$$

Now, Assumption 4.5'.(ii),(iv) and the Mean Value Theorem imply that, as $T \rightarrow \infty$, almost surely,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T [Q_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\beta) |_{\beta=\beta_t} - Q_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\beta_0)] = \frac{1}{\sqrt{R_T T}} \sum_{t=1}^T \left[D_{\beta} Q_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\tilde{\beta}_t) \right] \sqrt{R_T} (\beta_t - \beta_0),$$

with $\beta_t := [\beta_{t_1} \dots \beta_{t_M}]'$, and $\tilde{\beta}_t$ a random point on the ray that connects β_t and β_0 inside $\bar{B}_{\beta_0}(\eta)$. Due to Assumption 4.1'.(v), it follows that $\sup_{\bar{B}_{\beta_0}(\eta)} \|D_{\beta} Q_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\beta)\| < +\infty$, and then from Assumption 4.5'.(ii), (iii), (v), it follows that via van der Vaart (2000, Thm 18.14), the Cramer-Wold device, and similar considerations as in the proofs of West and McCracken (1998, Lemmata 4.1-2), that jointly with (27),

$$\frac{1}{\sqrt{R_T T}} \sum_{t=1}^T [D_{\beta} Q_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\beta)] \sqrt{R_T} (\beta_t - \beta_0) \rightsquigarrow \mathcal{G}_{S_*}(\beta), \text{ in } \ell^\infty(\bar{B}_{\beta_0}(\eta), \mathbb{R}^M) \quad (29)$$

where now $\mathcal{G}_{S_*}(\beta)$ is a Gaussian process with uniformly continuous sample paths in $\ell^\infty(\bar{B}_{\beta_0}(\eta))$, and covariance kernel

$$K_{\mathcal{G}_*}(\beta_1, \beta_2) = \varrho_* D_{\beta} Q_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\beta_1) \mathbf{H} V_h \mathbf{H}' D_{\beta} Q_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\beta_2)'$$

The definition of $\tilde{\beta}_t$ and (29) imply that jointly with (28),

$$\frac{1}{\sqrt{R_T T}} \sum_{t=1}^T \left[D_{\beta} Q_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\tilde{\beta}_t) \right] \sqrt{R_T} (\beta_t - \beta_0) \rightsquigarrow \mathcal{G}_{S_*}(\beta_0), \quad (30)$$

and, via considerations similar to those in the proofs of West and McCracken (1998, Lemmata 4.1-2), it is obtained that

$$\text{Cov}(\mathcal{G}_S(\beta_0), \mathcal{G}_{S_*}(\beta_0)) = \varrho \sum_{i=0}^{\infty} \kappa_i \text{Cov} \left(\frac{dL(\mathbf{U}_0 \boldsymbol{\lambda}_{\mathcal{F}}^*)}{dx} \mathbf{U}_0, D_{\beta} Q_{\boldsymbol{\lambda}_{\mathcal{F}}^*}(\beta_0) \mathbf{H} \mathbf{h}_i \right).$$

Then the result in (11) follows via the Continuous Mapping Theorem for $S_{\boldsymbol{\lambda}_{\mathcal{F}}^*} = \mathcal{G}_S(\beta_0) + \mathcal{G}_{S_*}(\beta_0)$. The result in (12) follows directly from Assumption 4.5'.(ii),(v) and the Uniform Law of Large Numbers of Birkhoff. The resulting Grammian matrix is positive definite due to Assumption 4.5'.(vi). Assumption 4.5.(i) and the absolute regularity part of Assumption 4.5'.(iii) imply the tightness in Assumption 4.5.(iii), via a modification of van der Vaart (2000, Cor 5.53), where the maximal inequality for identical and independent processes is substituted by the one in Dedecker and Louhichi (2002, Thm 3.3; see also Remark 1 on p. 146 for $b = 2$). The final part of Assumption 4.5.(iii) follows by Assumption 4.5'.(i) and the Dominated Convergence Theorem. ■

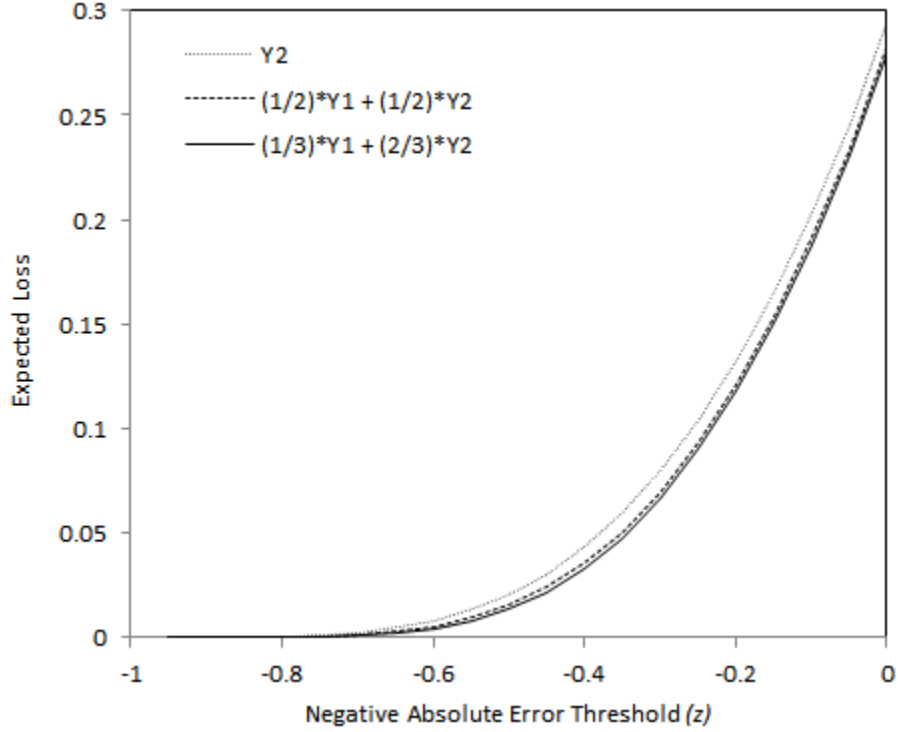


Figure 1: SCLSD Relations Between Three Forecast Combinations. The random variable X obeys a latent standard uniform distribution. Two independent forecasts ($M = 2$) are available: Y_1 is an independent standard uniform random variable; Y_2 is the mean of *two* of such random variables. For three distinct combinations of Y_1 and Y_2 , the figure shows the expected loss $\mathbb{E}_{\mathcal{F}}[L_z(\mathbf{U}\boldsymbol{\lambda})]$ as a function of the threshold level for negative absolute forecast error, $z \in \mathcal{Z}^- = [-1, 0]$. The three combinations are $(\lambda_1, \lambda_2) = (0, 1)$, $(\lambda_1, \lambda_2) = (\frac{1}{2}, \frac{1}{2})$, and $(\lambda_1, \lambda_2) = (\frac{1}{3}, \frac{2}{3})$. By Proposition 2.1, SCLSD occurs if and only if the optimal mixture reduces expected loss for every threshold level.

Table I: Simulated Properties of the Empirical Solution. X and Y_1 are independent standard uniform random variables; Y_2 is the mean of $n = 1, 2$ independent standard uniform random variables. A total of 10,000 independent random samples of $T = 10, 30, 100, 300, 1000$ paired observations of X, Y_1 and Y_2 are drawn. For every random sample, the latent CDF \mathcal{F} is estimated using the ECDF \mathcal{F}_T and the unconstrained optimization problem $\max_{\lambda \in \Lambda} G_{\mathcal{F}_T, \lambda}$ and the constrained optimization problem (7) are solved with the objective of minimizing the sample MSFE and the EWA as the benchmark. Given the optimal mixing weights $\lambda_{T,1}^*$ and $\lambda_{T,2}^*$, the population MSFE can be computed as $\left((\lambda_{T,1}^*)^2 + \frac{1}{n} (\lambda_{T,2}^*)^2 + 1 \right) \frac{1}{12}$. Shown are the 10th, 25th, 50th, 75th and 90th percentile breakpoints (P10, ..., P90) of the weight $\lambda_{T,1}^*$ and the MSFE.

Panel A: Two equivalent forecasts ($n = 1; \lambda_1^* = \frac{1}{2}; \text{MSFE} = \frac{1}{8}$)											
	T	w/o SD constraints					w/ SD constraints				
		P10	P25	P50	P75	P90	P10	P25	P50	P75	P90
Weight ($\lambda_{T,1}^*$)	10	0.164	0.320	0.495	0.682	0.857	0.286	0.500	0.500	0.500	0.748
	30	0.333	0.407	0.507	0.595	0.670	0.500	0.500	0.500	0.500	0.500
	100	0.389	0.444	0.499	0.555	0.598	0.500	0.500	0.500	0.500	0.500
	300	0.443	0.468	0.500	0.529	0.557	0.500	0.500	0.500	0.500	0.500
	1000	0.470	0.483	0.500	0.516	0.533	0.500	0.500	0.500	0.500	0.500
MSFE	10	0.125	0.126	0.130	0.141	0.157	0.125	0.125	0.125	0.128	0.148
	30	0.125	0.125	0.126	0.129	0.134	0.125	0.125	0.125	0.125	0.127
	100	0.125	0.125	0.126	0.126	0.128	0.125	0.125	0.125	0.125	0.125
	300	0.125	0.125	0.125	0.125	0.126	0.125	0.125	0.125	0.125	0.125
	1000	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
Panel B: Two non-equivalent forecasts ($n = 2; \lambda_1^* = \frac{1}{3}; \text{MSFE} = \frac{1}{9}$)											
	T	w/o SD constraints					w/ SD constraints				
		P10	P25	P50	P75	P90	P10	P25	P50	P75	P90
Weight ($\lambda_{T,1}^*$)	10	0.000	0.100	0.322	0.534	0.721	0.000	0.189	0.496	0.501	0.505
	30	0.131	0.229	0.337	0.463	0.552	0.176	0.342	0.491	0.498	0.502
	100	0.220	0.268	0.330	0.398	0.459	0.249	0.314	0.441	0.489	0.498
	300	0.264	0.297	0.329	0.365	0.395	0.277	0.316	0.365	0.431	0.473
	1000	0.295	0.315	0.334	0.353	0.370	0.300	0.321	0.343	0.370	0.399
MSFE	10	0.111	0.112	0.117	0.125	0.130	0.113	0.114	0.115	0.118	0.125
	30	0.111	0.111	0.113	0.116	0.120	0.111	0.113	0.114	0.115	0.115
	100	0.111	0.111	0.112	0.113	0.114	0.111	0.111	0.113	0.114	0.114
	300	0.111	0.111	0.111	0.112	0.112	0.111	0.111	0.111	0.112	0.114
	1000	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.112

Table II: VIX forecasting: Shown are the intercept, slope and R-squared of the Mincer-Zarnowitz regression of actual on forecasted values; the Diebold-Mariano statistic for MAFE of the optimized combination being equal to that of the Equal-Weighted Average (DM1); the Diebold-Mariano statistic for MSFE (DM2); the fraction of correct sign predictions; the (annualized) Information Ratio (IR) for a simple strategy of long (short) volatility when an increase (decrease) is predicted. Separate results are shown for the full sample (1994-2017), the Global Financial Crisis period (2007-2008) and the sample excluding the crisis period.

Period	Accuracy Measures										Average weights						
	c_0	c_1	R^2	DM1	DM2	Sign	IR	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7			
94-17	EWA	0.000	2.163	2.704		0.554		0.143	0.143	0.143	0.143	0.143	0.143	0.143			
	MAFE	0.000	1.250	2.747	3.554	2.665	-0.142	0.024	0.040	0.122	0.073	0.064	0.104	0.573			
	MAFE SD	0.000	1.520	3.001	4.920	3.528	0.027	0.045	0.075	0.177	0.095	0.084	0.126	0.399			
	MSFE	0.000	1.158	2.615	2.864	2.357	-0.009	0.036	0.083	0.165	0.121	0.026	0.082	0.487			
	MSFE SD	0.000	1.353	2.742	3.700	2.698	0.093	0.056	0.091	0.186	0.114	0.061	0.110	0.381			
GFC	EWA	0.003	1.961	3.079		0.554		0.143	0.143	0.143	0.143	0.143	0.143	0.143			
	MAFE	0.002	1.191	1.996	-0.709	-0.483	-0.613	0.116	0.214	0.160	0.009	0.050	0.013	0.438			
	MAFE SD	0.003	1.511	2.846	-0.254	0.386	-0.382	0.111	0.266	0.168	0.052	0.093	0.119	0.190			
	MSFE	0.002	0.935	1.967	-1.536	-0.388	-0.792	0.175	0.501	0.163	0.033	0.037	0.001	0.090			
	MSFE SD	0.003	1.118	2.213	-1.169	-0.176	-0.282	0.176	0.344	0.114	0.058	0.095	0.108	0.105			
94-17 ex GFC	EWA	0.000	2.209	2.657		0.554		0.143	0.143	0.143	0.143	0.143	0.143	0.143			
	MAFE	0.000	1.255	2.853	3.936	3.003	-0.086	0.016	0.024	0.118	0.079	0.065	0.112	0.586			
	MAFE SD	0.000	1.521	3.023	5.158	3.593	0.072	0.038	0.057	0.178	0.099	0.083	0.127	0.418			
	MSFE	0.000	1.197	2.724	3.577	2.744	0.016	0.023	0.044	0.165	0.129	0.025	0.090	0.524			
	MSFE SD	0.000	1.395	2.838	4.348	3.087	0.185	0.046	0.068	0.193	0.119	0.058	0.111	0.406			