Journal of Econometrics 242 (2024) 105767

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Better the devil you know: Improved forecasts from imperfect models[☆]

Dong Hwan Oh[a], Andrew J. Patton[b,*]

[a] *Federal Reserve Board, United States of America*
[b] *Duke University, United States of America*

ARTICLE INFO

JEL classification:
C53
C51
C58
C14

Keywords:
Model misspecification
Local maximum likelihood
Volatility forecasting

ABSTRACT

Many important economic decisions are based on a parametric forecasting model that is known to be good but imperfect. We propose methods to improve out-of-sample forecasts from a misspecified model by estimating its parameters using a form of local M estimation (thereby nesting local OLS and local MLE), drawing on information from a state variable that is correlated with the misspecification of the model. We theoretically consider the forecast environments in which our approach is likely to offer improvements over standard methods, and we find significant forecast improvements from applying the proposed method across four distinct empirical analyses including volatility forecasting, risk management, and yield curve forecasting.

1. Introduction

Many important economic decisions are based on a forecasting model that is known to be good but imperfect. Such a model may be retained for a variety of reasons: the model, and its flaws, may be well-studied and understood, unlike its possible replacement; there may be institutional impediments to adopting new models; the competitive environment may be such that it is not possible to switch to a new model in time for it to be of help. For example, central banks maintain a decision-making infrastructure around a given model or class of models, as do risk management departments at large financial institutions, and high-frequency trading algorithms have models physically built into the processing chips. In all of these cases, the model at the heart of these decisions is known to be good (else it would not have been embedded in the processes) however it is almost certainly also imperfect.

We propose a method to improve the out-of-sample forecasts from a misspecified model by estimating the parameters in a way that emphasizes epochs that are similar to the one in which the forecast is being made. Our approach exploits information from a state variable that is correlated with the misspecification of the model. For example, consider the case that the true data generating process (DGP) is a complicated nonlinear autoregressive process, and the model is a simple AR(1). Through experience, the forecast user may know that when the target variable is far from its average level the degree of mean-reversion tends to be stronger than when it is around its average value. This information can be used to "tilt" the AR parameter from its usual OLS estimate when the target variable is indeed further from its mean. We provide a structured approach for incorporating this useful information into the parameter estimate without altering the baseline model.

[☆] We thank Tim Bollerslev, Ana Galvao, Simone Manganelli, Mike McCracken, Rogier Quaedvlieg, Allan Timmermann (discussant) and seminar participants at the Australasian Finance & Banking Conference, European Central Bank, Deutsche Bundesbank, Federal Reserve Board, University of Connecticut, Monash University, Singapore Management University, and the SoFiE Seminar Series. The second author also thanks UNSW Sydney, where part of this work was completed, for their hospitality. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Federal Reserve Board.

* Corresponding author.
E-mail addresses: donghwan.oh@frb.gov (D.H. Oh), andrew.patton@duke.edu (A.J. Patton).

https://doi.org/10.1016/j.jeconom.2024.105767
Received 20 June 2022; Received in revised form 14 November 2023; Accepted 31 January 2024
Available online 23 May 2024

Formally, our method can be interpreted as a form of nonparametric estimation of the parameters of the baseline model. It is a folk theorem in economic forecasting that nonparametric methods perform poorly out-of-sample, as the increased estimation error overwhelms the improved fit of the model. We consider this canonical trade-off in a theoretical examination of our approach, and we identify two key aspects of the forecasting problem that influence the ability of our approach to improve upon standard methods. Firstly, if the baseline model is "too good," then there is little room for improvement and usual estimation approach will dominate. Fortunately or unfortunately, even popular forecasting models are inevitably misspecified, leaving open the possibility for improvement. Secondly, if the forecast user's experience does not yield an informative state variable, then our estimator will converge to the usual estimator's probability limit, but accompanied by greater estimation error. Widely-used models inevitably accumulate a lot of practical experience about their properties and pitfalls, and so it is commonly the case that an informative state variable is available.

We apply the proposed method to four economic forecasting problems. In the first two applications we consider volatility forecasting, either using the seminal GARCH model of Bollerslev (1986), or the popular alternative for models using high-frequency data, the HAR model of Corsi (2009), estimated by QML. Our third application considers joint forecasts of Value-at-Risk and Expected Shortfall (VaR and ES), and so the target functional is a (2×1) vector, estimated using M-estimation. Finally, we consider yield curve forecasts using the popular Diebold and Li (2006) model, estimated by OLS, with maturities ranging from three months to ten years. These four applications illustrate the variety of environments (target functionals, dimensionality, estimation methods), and we show that our proposed method provides statistically significant improvements over standard methods.

The estimation method proposed here is closely related to the local MLE of Tibshirani and Hastie (1987), Fan et al. (1998), and Fan et al. (2009), but unlike those approaches we do *not* modify the baseline model in an attempt to recover the DGP; instead we "tilt" the parameters of the model so that they better fit the current environment, and produce better forecasts.[1] Our approach is a mid-point between the fully parametric ML estimator and the fully nonparametric approach of Fan et al. (2009): we keep the model fully parametric, but we use nonparametric methods to optimally weight the observations used in the estimation window. In this sense, our approach is also similar to the "relevance-weighted ML" of Hu (1997), however we differ in that our weights arise from the chosen kernel and bandwidth, and we allow the bandwidth to go to zero, making this a nonparametric estimator.[2] Also related, but in a different context, Kristensen and Mele (2011) propose a method to obtain derivative prices by approximating the pricing error implied by a simple and well-known method (the Black–Scholes formula).

A well-known type of local estimation is rolling window estimation, which has been found to improve forecast performance in a variety of applications, particularly in the presence of structural breaks, see Pesaran et al. (2013), Inoue et al. (2017) and others. It is also similar to the use of exponential smoothing, see Brown (1956), Muth (1960), and Zumbach (2006), where more recent observations are given a higher weight in estimation than older observations. Both methods attempt to capture the fact that as the DGP evolves through time, the best-fitting approximating model will vary too. These methods correspond to using time as the state variable, and a one-sided rectangular or exponential kernel.[3] Related, Ang and Kristensen (2012) and Inoue et al. (2020) consider the estimation of factor models and GARCH models, respectively, with parameters that vary smoothly over time, though those papers focus on model estimation rather than prediction.

Dendramis et al. (2020) is perhaps the most closely-related paper to ours. That paper focuses on conditional mean forecasts made using ARMA models and estimated by OLS. The authors note that the gains they find are somewhat small and not always a significant improvement over their benchmark AR(1) model. This is in contrast with the variety of target variables, functionals, and estimation methods that we consider, and the robust and strongly significant gains in forecast performance that we find empirically. Further, we theoretically analyze the bias–variance trade-off present in a local estimation framework, and obtain predictions for when such a method is likely to work well in practice.

Our approach is also related to work on bringing outside information to bear on a forecasting problem. Manganelli (2009) considers the case that the forecaster has a "default decision" and provides a structured method for tilting a model-based forecast towards the default decision. Giacomini and Ragusa (2014) and Pettenuzzo et al. (2014) provide methods for adjusting model-based forecasts so that they satisfy constraints suggested by economic theory. The approach proposed in this paper requires less of the forecaster: no default decision and no economic theory, only a variable that is thought to be related to the degree of model misspecification.

Exploiting the expertise of the forecast user to identify a state variable to improve the forecasts obtained from a baseline model is also related to professional forecasters' use of both statistical models and expert judgment. Numerous studies, see Ang et al. (2007) and Faust and Wright (2009) for example, have found that professional forecasters regularly outperform standard model-based forecasts. Our tilting of the model parameters may be interpreted as a form a "structured" expert judgment, and the generally superior performance of our proposed method is consistent with this literature.

The remainder of the paper is structured as follows. In Section 2 we present our estimator and theoretically consider the bias–variance trade-off for local and non-local estimation methods in out-of-sample forecasting. In Section 3 we apply our estimator to four economic forecasting problems and Section 4 concludes. A supplemental appendix contains additional details and results.

[1] More specifically, we follow Fan et al. (1998) in the kernel-weighting of the likelihoods, but we do not take an expansion of the functional of interest in the state variable. Instead, we retain the specification of that functional as given by the baseline model.

[2] Blasques et al. (2016) also consider a weighted ML method, for applications where the vector of dependent variables can be separated into those of particular interest and the rest, and in estimation the likelihood of the former is overweighted relative to the latter.

[3] Theoretically, the interpretation of the local estimator differs in these applications: with a stochastic state variable one may still assume stationarity, while when using time as the state variable one must instead consider heterogeneity in the DGP, usually in the form of smoothly evolving parameters. Empirically, either form of state variable is equally easy to handle, and we consider both in our empirical applications.

2. Local estimation and out-of-sample forecasting

We consider a target variable Y_{t+1}, and target functional $g_t \in \mathcal{G}$. For example, g_t could be the mean, variance, median, a quantile, etc. It may also, with some changes in notation and methods, be a predictive density, though we will focus on point forecasting. The target functional may also be a vector, e.g. if Y_{t+1} is a vector and g_t is its mean, or if Y_{t+1} is a scalar and g_t is the vector containing the Value-at-Risk and Expected Shortfall. The forecaster's information set is $\mathcal{F}_t$, and naturally g_t is $\mathcal{F}_t$-measurable. We focus on one-step-ahead forecasts, but all the results below can be extended to general h- step-ahead forecasts, for $h < \infty$.

Let L be a loss function (scoring rule) that elicits the desired target functional, i.e., that

$$g_t^{\dagger} = \arg\min_{g\in\mathcal{G}} \mathbb{E}\left[L\left(Y_{t+1}, g\right)|\mathcal{F}_t\right] \tag{1}$$

For example, if the target functional is the mean, then L can be the squared forecast error.[4] The baseline model is a parametric model for the target functional, $g_t(\theta)$, and we assume the parameter of the model is obtained via M-estimation minimizing the same loss function. Matching the estimation and evaluation loss functions is intuitive and can lead to improved forecasts, see Granger (1969), Weiss (1996), and Hansen and Dumitrescu (2022) for example.[5]

$$\hat{\theta}_T = \arg\min_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} L\left(Y_t, g_{t-1}(\theta)\right) \tag{2}$$

where $\theta \in \Theta \subseteq \mathbb{R}^p$. We assume that the sample runs from $t = 0, 1, \ldots, T$, yielding T observations for estimation. Under standard conditions the usual estimator converges at rate $\sqrt{T}$ to a well-defined probability limit, $\hat{\theta}^*$, and has a Normal asymptotic distribution:

$$\hat{\theta}^* \equiv \arg\min_{\theta\in\Theta} \mathbb{E}\left[L\left(Y_{t+1}, g_t(\theta)\right)\right] \tag{3}$$

$$\sqrt{T}\left(\hat{\theta}_T - \hat{\theta}^*\right) \xrightarrow{D} N(0, \Sigma) \tag{4}$$

2.1. Incorporating information from a state variable

Denote the forecaster's state variable as S_t, with support $\mathcal{S} \subset \mathbb{R}^d$. This variable must be observable at all t (i.e., is $\mathcal{F}_t$-measurable), and may or may not be one of the variables in the baseline model. We consider an estimator defined by:

$$\tilde{\theta}_{h,T}(s) = \arg\min_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} L\left(Y_t, g_{t-1}(\theta)\right) K\left(s - S_{t-1}; h_T\right), \text{ for } s \in \text{Int}(\mathcal{S}) \tag{5}$$

where K is the kernel function, h_T is a bandwidth parameter that shrinks with the sample size, s is some prespecified value of the state variable, and $\text{Int}(\mathcal{S})$ is the interior of the support of the state variable, $\mathcal{S}$. Under a variety of regularity conditions, the limit of this estimator is:

$$\tilde{\theta}^*(s) \equiv \arg\min_{\theta\in\Theta} \mathbb{E}\left[L\left(Y_{t+1}, g_t(\theta)\right)|S_t = s\right] \tag{6}$$

With the bandwidth shrinking at an appropriate rate, which differs depending on assumptions about smoothness and temporal dependence,[6] the rate of convergence for the estimator is $T^{1/2-\gamma}$ for some $\gamma \in (0, 1/2)$:

$$T^{1/2-\gamma}\left(\tilde{\theta}_{h,T}(s) - \tilde{\theta}^*(s)\right) = \mathcal{O}_p(1) \quad \forall\ s \in \text{Int}(\mathcal{S}) \tag{7}$$

For the purposes of our analysis below, we require only that the estimator is consistent (so $\gamma < 1/2$) but converges more slowly than the parametric rate ($\gamma > 0$). Naturally, in applied work one would like to find the local estimator with the fastest rate of convergence, and in our applications we use cross-validation to choose the bandwidth that minimizes average loss.[7]

2.2. The special case of correct specification

Consider the special case that the baseline model is correctly specified and point identified (and so $\hat{\theta}^*$ is unique) for the target functional. This implies

$$\exists!\ \hat{\theta}^* \in \Theta \text{ s.t. } g_t^{\dagger} = g_t(\hat{\theta}^*) \text{ a.s. } \forall\ t \tag{8}$$

[4] As discussed in Gneiting (2011) and Patton (2020), in many cases there are an infinite number of loss functions that elicit a given functional.

[5] Hansen and Dumitrescu (2022) show that in some applications there may be gains from using a different loss function for estimation than evaluation, but only if the two loss functions are "coherent." We discuss this further in Supplemental Appendix SA.2.

[6] See Härdle and Tsybakov (1997) and Härdle et al. (1998) for results on the asymptotic distribution of local polynomial and local linear estimation for time series, and see Fan and Yao (2003), Chapter 6) for a survey of results on nonparametric estimation for time series.

[7] We focus on the case of a stochastic state variable here but the results below go through when conditioning instead on time, as the fundamental trade-off between a better local approximation and greater estimation error remains the same. The rate of convergence of the local estimator when using time as the state variable can again be shown to be $T^{1/2-\gamma}$ for some $\gamma \in (0, 1/2)$ under a variety of conditions, see Ang and Kristensen (2012) for example.

Now consider the population local estimator using today's value of the state variable

$$\tilde{\theta}^*\left(S_t\right) \equiv \arg\min_{\theta\in\Theta}\ \mathbb{E}\left[L\left(Y_{t+1}, g_t\left(\theta\right)\right)|S_t\right] \tag{9}$$

Since local estimation nests non-local estimation, we have

$$\mathbb{E}\left[L\left(Y_{t+1}, g_t\left(\tilde{\theta}^*\left(S_t\right)\right)\right)|S_t\right] \leq \mathbb{E}\left[L\left(Y_{t+1}, g_t\left(\hat{\theta}^*\right)\right)|S_t\right] \quad \text{a.s. } \forall\ t \tag{10}$$

Since $g_t^\dagger = \arg\min_{g\in\mathcal{G}}\ \mathbb{E}\left[L\left(Y_{t+1}, g\right)|\mathcal{F}_t\right]$, we also have

$$\mathbb{E}\left[L\left(Y_{t+1}, g_t^\dagger\right)|\mathcal{F}_t\right] \leq \mathbb{E}\left[L\left(Y_{t+1}, g_t\left(\tilde{\theta}^*\left(S_t\right)\right)\right)|\mathcal{F}_t\right] \tag{11}$$

and by correct specification we have

$$\mathbb{E}\left[L\left(Y_{t+1}, g_t^\dagger\right)|\mathcal{F}_t\right] = \mathbb{E}\left[L\left(Y_{t+1}, g_t(\hat{\theta}^*)\right)\Big|\mathcal{F}_t\right] \quad \text{a.s. } \forall\ t,\ \forall\ \theta\in\Theta \tag{12}$$

Since $S_t \in \mathcal{F}_t$, we can apply the law of iterated expectations (LIE) to Eqs. (11)–(12) and combine with Eq. (10) to obtain

$$\mathbb{E}\left[L\left(Y_{t+1}, g_t(\hat{\theta}^*)\right)\Big|S_t\right] \leq \mathbb{E}\left[L\left(Y_{t+1}, g_t\left(\tilde{\theta}^*\left(S_t\right)\right)\right)\Big|S_t\right] \leq \mathbb{E}\left[L\left(Y_{t+1}, g_t(\hat{\theta}^*)\right)\Big|S_t\right] \quad \text{a.s. } \forall\ t$$

Thus we have $\mathbb{E}\left[L\left(Y_{t+1}, g_t(\hat{\theta}^*)\right)\Big|S_t\right] = \mathbb{E}\left[L\left(Y_{t+1}, g_t\left(\tilde{\theta}^*\left(S_t\right)\right)\right)\Big|S_t\right]$ a.s. $\forall\ t$. By the point-identification assumption, we then know that $\tilde{\theta}^*\left(S_t\right) = \hat{\theta}^*$. Noting that this must be true (a.s.) for all t, this implies that $\tilde{\theta}^*(s)$ is flat in s. That is, the local M estimator reduces to the usual M estimator when the baseline model is correctly specified.

2.3. Out-of-sample forecasting and a bias–variance trade-off

We now consider out-of-sample (OOS) forecast accuracy using the local estimator and the usual, non-local, estimator. We obtain a form of bias–variance trade-off, which illuminates the conditions under which the local estimator is likely to outperform the usual estimator. As is common in the literature, we focus here on unconditional forecast accuracy, but it is also possible to look at the difference in forecast performance conditional on the value of a state variable, for example by using the methods of Li et al. (2022), or as a function of time, as in Giacomini and Rossi (2010) and Richter and Smetanina (2020). In our empirical applications we consider both unconditional and conditional performance.

By local estimation optimization, we have

$$\mathbb{E}\left[L\left(Y_{t+1}, g_t\left(\tilde{\theta}^*\left(S_t\right)\right)\right)|S_t\right] \leq \mathbb{E}\left[L\left(Y_{t+1}, g_t\left(\theta\right)\right)|S_t\right] \text{ a.s. } \forall\ t,\ \forall\ \theta\in\Theta \tag{13}$$

and by evaluating the right-hand side at the non-local estimator and invoking the LIE we obtain

$$\mathbb{E}\left[L\left(Y_{t+1}, g_t\left(\tilde{\theta}^*\left(S_t\right)\right)\right)\right] \leq \mathbb{E}\left[L\left(Y_{t+1}, g_t(\hat{\theta}^*)\right)\right] \tag{14}$$

This shows that the out-of-sample average loss from the local estimator will be weakly smaller than that from the usual estimator in population. Note that this is true even though OOS performance is computed using *non* -weighted losses, that is, the kernel function used in the local estimation objective function does not appear. The gains accrue because the local estimator can vary with the realized value of the state variable, while the usual estimator is fixed. As shown in the previous section, when the model is correctly specified we have $\tilde{\theta}^*(s) = \hat{\theta}^*\ \forall\ s$ and so local and non-local estimators are identical and yield identical expected loss.

Next we consider the variance of the estimators, and the deleterious impact that estimation error has on expected loss. It is this aspect that often makes forecasts from nonparametric models worse than those from parametric models. We do so using a second-order Taylor series expansion of the time $T+1$ expected loss incurred using the estimated parameter, centered on the limiting parameter. We assume that the unconditional expected loss, $\mathbb{E}\left[L\left(Y_{t+1}, \cdot\right)\right]$, and conditional expected loss, $\mathbb{E}\left[L\left(Y_{t+1}, \cdot\right)\Big|S_t\right]$, are, for all $\theta\in\Theta$, twice differentiable.[8,9] For ease of presentation we assume that $\dim(\theta) = 1$, which can easily be relaxed.

Consider firstly the expected loss from non-local estimation. By the $\sqrt{T}$-consistency of $\hat{\theta}_T$ and Taylor's theorem, we have

$$\begin{aligned}\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\hat{\theta}_T\right)\right)\right] &= \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\hat{\theta}^*\right)\right)\right] + \frac{\partial\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\hat{\theta}^*\right)\right)\right]}{\partial\theta}\left(\hat{\theta}_T - \hat{\theta}^*\right) \\ &\quad + \frac{1}{2}\frac{\partial^2\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\hat{\theta}^*\right)\right)\right]}{\partial\theta^2}\left(\hat{\theta}_T - \hat{\theta}^*\right)^2 + o_p\left(T^{-1}\right)\end{aligned} \tag{15}$$

[8] Note that this assumption can hold even when the loss function is not differentiable, as in our use of the "FZ0" loss function (Fissler and Ziegel, 2016; Patton et al., 2019) in our VaR-ES application, if the target variable is continuously distributed.

[9] Twice differentiability of the objective function is a common regularity condition in estimation, see Newey and McFadden (1994) for example. Primitive conditions that guarantee this holds depend on the specific loss function, L, forecasting model, g_t, and conditional distribution of the target variable, $Y_{t+1}|\mathcal{F}_t$. This condition holds under a range of different assumptions on the model and data generating process: see White (2001) for sufficient conditions that ensure these hold for linear regression (as in our yield curve applications); see Newey and McFadden (1994) and White (1994) for sufficient conditions for QMLE and M estimation (as in our GARCH and HAR applications); see Patton et al. (2019) for sufficient conditions for M estimation using the "FZ0" loss function (as in our VaR-ES application).

The first-order term is zero since $\partial \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\hat{\theta}^*\right)\right)\right] / \partial \theta = 0$ by the definition of $\hat{\theta}^*$. The second-order term is a Hessian-like term:

$$\hat{H}^* \equiv \frac{1}{2} \frac{\partial^2 \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\hat{\theta}^*\right)\right)\right]}{\partial \theta^2} \tag{16}$$

which is positive in standard estimation problems. Thus the second-order term is positive and vanishing at rate T^{-1}. Next consider expected loss from local estimation. By the $T^{1/2-\gamma}$ consistency of $\tilde{\theta}_T(\cdot)$ we have

$$\begin{aligned}\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}_{h,T}\left(S_T\right)\right)\right)\right] = {} & \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*\left(S_T\right)\right)\right)\right] \\ & + \frac{\partial \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*\left(S_T\right)\right)\right)\right]}{\partial \theta}\left(\tilde{\theta}_{h,T}\left(S_T\right) - \tilde{\theta}^*\left(S_T\right)\right) \\ & + \frac{1}{2} \frac{\partial^2 \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*\left(S_T\right)\right)\right)\right]}{\partial \theta^2}\left(\tilde{\theta}_{h,T}\left(S_T\right) - \tilde{\theta}^*\left(S_T\right)\right)^2 + o_p\left(T^{-1+2\gamma}\right)\end{aligned} \tag{17}$$

The first-order term is zero since

$$\begin{aligned}\frac{\partial}{\partial \theta} \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*\left(S_T\right)\right)\right)\right] &= \frac{\partial}{\partial \theta} \mathbb{E}\left[\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*\left(S_T\right)\right)\right) \middle| S_T\right]\right] \\ &= \mathbb{E}\left[\frac{\partial}{\partial \theta} \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*\left(S_T\right)\right)\right) \middle| S_T\right]\right] = 0\end{aligned} \tag{18}$$

The first equality holds by the law of iterated expectations, the second holds as we can interchange differentiation and (unconditional) expectation under the twice differentiability assumption, and the third equality holds since $\partial \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*\left(S_T\right)\right)\right) \middle| S_T\right] / \partial \theta = 0$ from the definition of $\tilde{\theta}^*\left(S_T\right)$. The second-order term involves an expected Hessian-like term:

$$\tilde{H}^*\left(S_T\right) \equiv \frac{1}{2} \frac{\partial^2 \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*\left(S_T\right)\right)\right)\right]}{\partial \theta^2} = \frac{1}{2} \mathbb{E}\left[\frac{\partial^2 \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*\left(S_T\right)\right)\right) \middle| S_T\right]}{\partial \theta^2}\right] \tag{19}$$

which is positive in standard estimation problems.

Finally, consider the difference between the OOS losses using the above two expansions:

$$\begin{aligned}& \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}_{h,T}\left(S_T\right)\right)\right) - L\left(Y_{T+1}, g_T\left(\hat{\theta}_T\right)\right)\right] \\ = {} & \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*\left(S_T\right)\right)\right) - L\left(Y_{T+1}, g_T\left(\hat{\theta}^*\right)\right)\right] + \mathcal{O}_p\left(T^{-1+2\gamma}\right)\end{aligned} \tag{20}$$

The first term on the right-hand side is non-positive, as the local estimator has weakly smaller expected loss than the usual estimator when both are evaluated at population parameters. The second term is dominated by the magnitude of the estimation error in the local estimator, and is positive and vanishing at rate $T^{-1+2\gamma}$. Using an optimal bandwidth will make γ as small as possible, reducing the magnitude of the estimation error in the local estimator. Since this term is positive, it increases the expected loss from using estimated parameters, and we observe the usual trade-off in forecasting: a more flexible model leads to improved fit, at a cost of increased estimation error. Whether one of these terms outweighs the other depends on features specific to each application, and we discuss these next.

2.4. Empirical predictions from the theoretical analysis

Firstly, consider the case that the baseline model is correctly specified. In that case Section 2.2 showed that $\tilde{\theta}^*(s) = \hat{\theta}^* \; \forall \; s$, and we have

$$\mathbb{E}[\underbrace{L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*\left(S_T\right)\right)\right)}_{\text{local estimator loss}} - \underbrace{L\left(Y_{T+1}, g_T(\hat{\theta}^*)\right)}_{\text{non-local estimator loss}}] = 0 \tag{21}$$

In this case, there is no improvement in the fit from using local estimation, and increased estimation error causes local estimation to have worse OOS performance. More generally, when the baseline model is "very good" the scope for an improvement in fit is reduced, and the possibility that any such improvements are more than offset by increased estimation error is increased.

Secondly, consider the case that the state variable contains no information about variation in the fit of the misspecified model. We quantify this by considering the population first-order conditions (FOCs) for the estimation methods. If the scores of the usual, non-local, estimator are *mean independent* of the state variable S_t, i.e.,

$$\mathbb{E}\left[\frac{\partial L\left(Y_{t+1}, g_t(\hat{\theta}^*)\right)}{\partial \theta} \middle| S_t\right] = \mathbb{E}\left[\frac{\partial L\left(Y_{t+1}, g_t(\hat{\theta}^*)\right)}{\partial \theta}\right] \tag{22}$$

then the local estimation's FOC is satisfied when $\tilde{\theta}^*\left(S_t\right) = \hat{\theta}^*$, since the RHS of the above equation equals zero by the FOC of the usual estimator. Thus a worthless state variable leads to $\tilde{\theta}^*(s)$ being flat in s. This is the same outcome as in the correctly specified case, although from a different source, namely the use of a poor state variable.[10] Since $\tilde{\theta}^*\left(S_t\right) = \hat{\theta}^*$ in this case, there is

[10] In the correctly-specified case, the scores are a MDS with respect to $\mathcal{F}_t$, and since $S_t \in \mathcal{F}_t$ the LIE implies $E\left[\partial L\left(Y_{t+1}, g_t\left(\hat{\theta}^*\right)\right) / \partial \theta \middle| S_t\right] = 0$ for any choice of S_t, implying that in this case there are *no* useful state variables.

obviously no improvement in the fit from using local estimation, and the estimation error term discussed in the previous section causes local estimation to have worse OOS performance. More generally, when the state variable is only weakly informative about model misspecification the gains from local estimation are lower, and the possibility that any such gains are more than offset by increased estimation error is increased.

2.5. A stylized example

To illustrate the above ideas, consider a nonlinear AR(1) process as the DGP and a standard AR(1) as the baseline model. Concretely, we use a stationary copula-based Markov process (see, e.g., Chen and Fan (2006) and Beare (2010)), with standard Normal marginal distributions and a Clayton copula linking adjacent realizations:

$$\left(Y_t, Y_{t-1}\right) = C_{Clayton}\left(\Phi, \Phi; \kappa\right) \tag{23}$$

where Φ is a standard Normal CDF, and κ is the parameter of the Clayton copula. We set $\kappa = 5$ which implies first-order autocorrelation of about 0.85, and consider an estimation sample of $T = 1000$. The conditional mean of Y_t given Y_{t-1} is nonlinear in Y_{t-1} for this process, and in the upper panel of Fig. 1 we see that it is increasing and concave. The upper panel of Fig. 1 also shows the fitted linear AR(1) prediction obtained by OLS.

If we use Y_{t-1} as the state variable for local OLS estimation, which was one of the key state variables considered in Dendramis et al. (2020), then in this example the local estimator asymptotically recovers the true conditional expectation function, since the truth is a nonlinear AR(1). That is, in this example local estimation completely fixes the misspecification of the linear AR(1) model. This estimator is denoted "Local OLS 1," and the upper panel of Fig. 1 confirms that this estimator closely tracks the true conditional expectation function.[11] We also consider a local estimator using the second lag of the dependent variable, which is correlated with the ideal state variable but imperfect. The resulting estimated conditional expectation function is approximately correct for $Y_{t-1} < 0$, where first-order dependence is particularly strong for this process, but is noticeably incorrect for $Y_{t-1} > 0$, where dependence is weaker and the state variable is worse.

The lower panel of Fig. 1 presents the out-of-sample RMSE for the two local estimators across a range of bandwidth parameters. For the optimal choice of bandwidth ($h = 0.41$) the RMSE of first local estimator is almost equal to the RMSE of the optimal forecast, which of course represents the lower bound on RMSE. The RMSE of the second local estimator is greater than that of the first, consistent with this estimator using a worse state variable, and it is below the usual OLS estimator's RMSE for all but the smallest choices of bandwidth. (The optimal bandwidth is $h = 0.62$.) As the bandwidth grows the two local estimators generate RMSE that converges to that of the OLS estimator, as in that case the local estimators reduce to the OLS estimator.

3. Empirical applications

We consider our new estimation method in four different empirical applications. Firstly, we consider the widely-used GARCH model of Bollerslev (1986). In this application the target variable (returns) and the target functional (conditional variance) are both scalars, and the model is estimated using quasi maximum likelihood (QML). In our second application we consider a popular high-frequency successor to the GARCH model, namely the HAR model of Corsi (2009). In this application the target variable functional is again a scalar, and estimation is again done via QML. Our third application considers forecasts of Value-at-Risk and Expected Shortfall (VaR and ES), and so the target functional is a (2×1) vector, and the model is estimated using M-estimation. Finally, we consider yield curve forecasts using the popular "dynamic Nelson–Siegel" model of Diebold and Li (2006). In this case the target variable is a (12×1) vector of yields for bonds with maturities ranging from three months to ten years and the target functional is the conditional mean of that vector, estimated using OLS. These four applications illustrate the variety of environments (target functionals, dimensionality, estimation methods), and we show that our proposed method provides statistically significant improvements over standard methods.

Across all four applications, for stochastic state variables we use a Gaussian kernel:

$$K_G\left(x; h\right) = \exp\left\{-\frac{x^2}{2h^2}\right\},\ x \in \mathbb{R},\ h > 0 \tag{24}$$

We consider values for the bandwidth, h, in the range $0.01\sigma_S$ to $4\sigma_S$, where σ_S is the standard deviation of the state variable. A small value of h makes the model parameters more "local," but also decreases their precision since the effective sample size is smaller, and as h diverges the local estimator approaches the benchmark non-local method. We also consider an infinite bandwidth by comparing the average loss from the best finite bandwidth with that from the non-local method. When using time as a state variable we use a one-sided exponential kernel with bandwidth parameter λ and window length m:

$$K_E\left(j; \lambda\right) = \lambda^j\left(1-\lambda\right)/\left(1-\lambda^m\right)\mathbf{1}\left\{j < m\right\},\ j \in 0, 1, 2, \ldots \tag{25}$$

We consider values for λ ranging from 0.98 to 0.9999. Smaller values of λ imply that older data are given less weight in estimation, making the model parameters more local (in time) but subject to greater estimation error. As $\lambda \to 1$ the weight function becomes

[11] For each of the "local OLS" estimated conditional expectation functions in the upper panel we use the bandwidths identified as optimal according to the lower panel of Fig. 1.

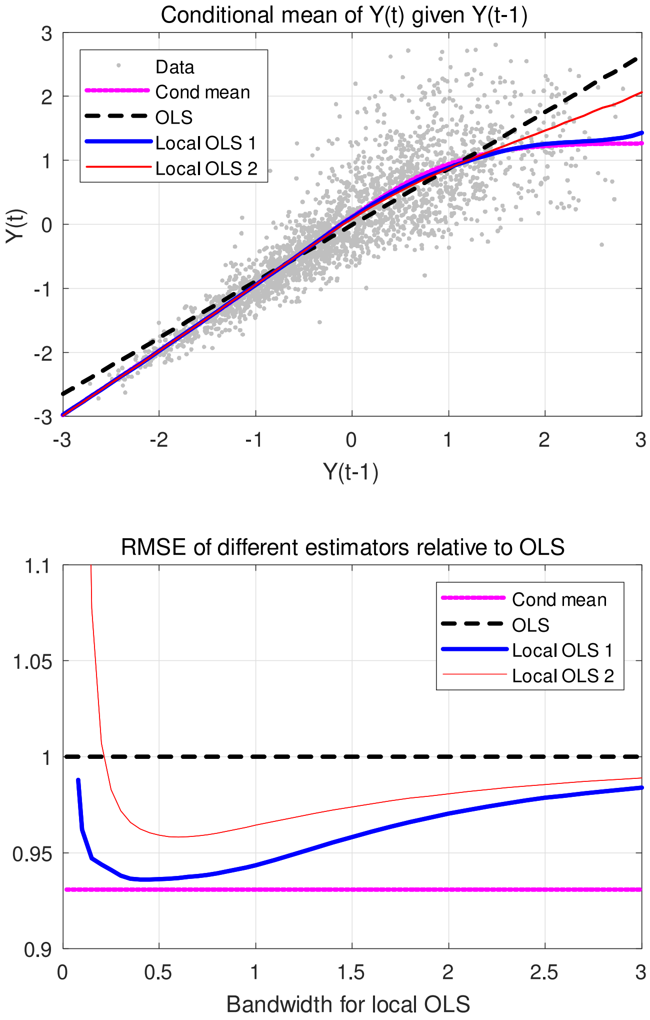

Fig. 1. The upper panel presents the expected value of Y_t given Y_{t-1} according to the DGP in Eq. (23), and estimates of this using a linear AR(1) estimated by OLS and local OLS with two different state variables: Y_{t-1} and Y_{t-2}. The lower panel presents the RMSE of the different estimators as a function of the local OLS bandwidth parameters.

flat and the local estimator approaches the benchmark non-local estimator. We consider the limiting case of $\lambda = 1$ by comparing the smallest average loss from a bandwidth less than 1 with the loss from the non-local method.

To select the optimal bandwidth parameter(s) for each state variable, we split the estimation sample into a "training sample" (the first half) for estimation of the model parameters with a variety of bandwidths, and a "validation sample" (the second half) to select the optimal bandwidth parameter(s).[12] We then use the selected bandwidth parameter when evaluating the model in the out-of-sample (OOS) period, eliminating look-ahead bias in both the model parameters and bandwidth parameters. Model parameters, for both the local and benchmark (non-local) models, are re-estimated daily throughout the OOS period using an expanding window of data, and bandwidth parameters for the local models are kept fixed at their optimized value from the validation sample.[13]

In all applications we consider four stochastic state variables, motivated by our applications to volatility or risk forecasting and yield curve forecasting. We consider two measures of volatility: 5-minute realized volatility (RV) on the S&P 500 index,[14] and

[12] For the bandwidth h we use a coarse grid of width 0.1 from $0.1\sigma_S$ to $4\sigma_S$ to find an approximate solution and then consider a finer grid of width 0.01 in an interval ± 0.1 from the approximate solution. For the bandwidth λ we consider a grid of width 0.0025 from 0.98 to 1, but we replace 1 with 0.999, 0.9995 and 0.9999.

[13] An earlier version of this paper considered non-local models estimated on rolling windows of 250, 500 and 1000 days of data. For all but the 1-day yield curve application we found very similar results to those reported here; non-local estimation was significantly out-performed by local estimation. For the 1-day yield curve application we found that the best two models were "non-local" using estimation windows of 250 and 500 days. However, using such short estimation windows also corresponds to a crude form of local estimation, one that is well known in the literature.

[14] This data is taken from the Oxford-Man Realised Library.

VIX, a measure of S&P 500 index volatility extracted from options prices. We also consider two measures derived from the yield curve: the Federal Funds Rate (FFR) and the difference between 10-year and 2-year government bond yields (denoted 10Y-2Y), representing measures of the "level" and "slope" of the yield curve. To mitigate skewness we use the natural logarithm of the two volatility measures. We also consider time as a state variable, and four bivariate state variables comprised of time and each of the four stochastic state variables, leading to a total of nine possible state variables. As the kernel for the bivariate state variables we use the product of the univariate kernel for each of the variables. Supplemental Appendix SA.1 provides additional details on the implementation of the local methods.

In our main analyses, we compare the various estimation methods in each application using OOS average loss. Importantly, OOS losses are *unweighted*, and so the local estimator has no inherent advantage; any forecast performance improvements are attributable to a favorable bias–variance trade-off relative to the benchmark method, in the spirit of the analysis in Section 2. We use Giacomini and White (2006) (GW) tests to compare each method to the benchmark non-local method, and we estimate the set of best methods using the model confidence set (MCS) of Hansen et al. (2011).[15] Digging deeper into the comparison of the competing methods, in Section 3.5 we consider *conditional* analyses of forecast performance, investigating whether relative performance varies with the state variable.

3.1. GARCH forecasts

The GARCH model of Bollerslev (1986) is a very popular model for forecasting asset return volatility, and in a variety of applications, and against a variety of alternatives (see Hansen and Lunde (2005)), it has proven hard to beat.[16] Assuming the conditional mean is zero, the GARCH model for the conditional volatility of asset return Y_t is:

$$\begin{aligned} Y_t &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= \omega + \beta \sigma_{t-1}^2 + \alpha Y_{t-1}^2 \end{aligned} \tag{26}$$

The benchmark method estimates the model parameters using QML, which is equivalent to minimizing the in-sample average QLIKE loss function:

$$L\left(Y_t^2, \sigma_t^2\right) = \frac{Y_t^2}{\sigma_t^2} - \log \frac{Y_t^2}{\sigma_t^2} - 1 \tag{27}$$

For this analysis we use daily returns on the S&P 500 index over the period January 2000 to June 2021, a total of $T = 5349$ observations. We use the period 2000–2010 (2737 observations) as the estimation sample, which is then further split into two to select the bandwidth parameters, and the remainder (2612 observations) as the out-of-sample period.

Table 1 presents the out-of-sample performance of the GARCH(1,1) model estimated using a variety of methods. The rows of this table are ordered by average OOS QLIKE loss, reported in the third-last column. The local method with the best performance in the validation sample (the second half of the in-sample period) is marked in the first column with $*$. The last two columns report Giacomini-White t-statistics of each model relative to the benchmark non-local model, and an indicator (✓or ×) for whether a given method is included in the 95% model confidence set.

We observe that the benchmark method, which uses non-local QML and an expanding estimation window, is ranked *last* in this set of estimation methods. Every local method has significantly lower OOS loss than the benchmark method, according to the GW test, with GW t-statistics all being less than −3. The local method with the best performance in the validation sample uses time and RV as state variables, and it turns out to also have the lowest average loss in the OOS period. Comparing the benchmark method with the local method selected using the validation sample we obtain a GW t-statistic of −12.3, strong evidence that the local method out-performs the non-local benchmark. When we consider this set of estimation methods as a whole, we find only one method is included in the model confidence set: local QML using time and RV as state variables. This small MCS indicates a high degree of precision in identifying the best-performing method.

To better understand the source of the improvement in performance of the best local method, Fig. 2 presents the local QML estimates of the GARCH parameters when RV ranges over its support, and compares them with the usual, constant, QML estimates of these parameters. To facilitate interpretation we look at three functions of these parameters: the model-implied average volatility ($\sqrt{\omega/(1-\alpha-\beta)}$), reaction of volatility to news (α), and persistence of volatility ($\alpha+\beta$). We see that the local QML estimate of the level of volatility is increasing in RV, consistent with RV providing useful information about future volatility. In the second panel we see that the reaction to news from local QML is generally lower than from non-local QML, and it is highest when RV is around 40, indicating that it is these times where the squared return is most informative about future volatility. We also observe a drop in the persistence of volatility when RV is high; above about 35. This is consistent with some successful extensions of the GARCH model, e.g., where volatility is modeled as having a fast- and a slow-moving component (see Engle and Lee (1999) and Christoffersen et al. (2008)) with sharp increases in volatility being attributable to the less-persistent component, or, related, where volatility is modeled as having a jump and a continuous component (Andersen et al. (2007)), with the jump component found to be less persistent.

[15] We use Newey–West standard errors with ten lags for the GW test, and we use the stationary bootstrap with an average block length of ten for the MCS.

[16] There are many papers that have built on the original GARCH model, and we do not attempt to conduct a horserace of volatility models here. Rather we illustrate how our method improves upon the seminal GARCH model, and, aside from one exception discussed at the end of this section, leave applying the method to extensions for future research.

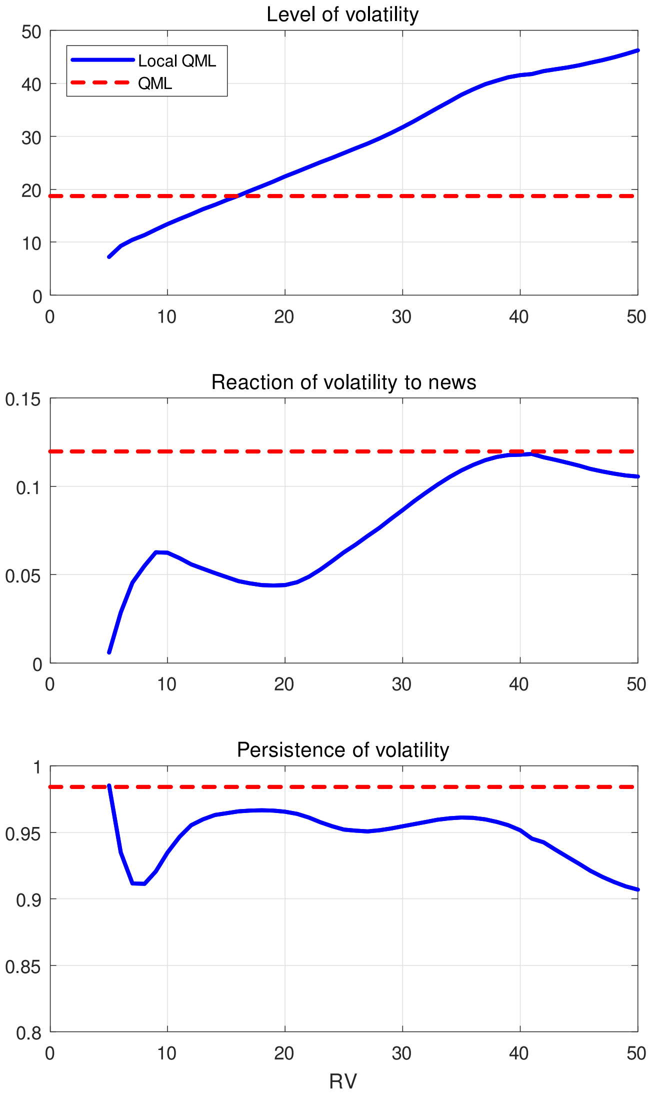

Fig. 2. This plot shows the local QML estimates of transformations of the GARCH(1,1) parameters (ω, β, α) as a function of realized volatility (RV). Also shown are the (non-local) QML parameter estimates. The upper, middle and lower panels plot $\sqrt{\omega/(1-\alpha-\beta)}$, α, and $(\alpha+\beta)$ respectively.

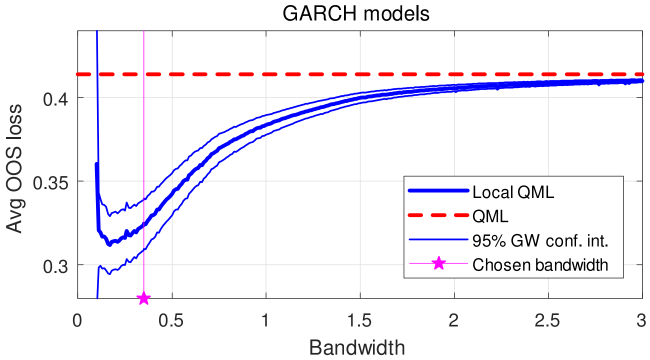

Fig. 3. This plot shows the average out-of-sample (OOS) loss using local QML estimates of a GARCH model as a function of the bandwidth. The state variable is realized variance. The dashed line is the (non-local) QML average loss, and the thin solid lines show the 95% GW confidence interval on the difference between local and non-local average loss. The starred vertical line shows the bandwidth chosen based on a validation sample separate from the OOS period.

Table 1
Out-of-sample forecast performance for GARCH(1,1) models.

Rank	*Method details*		*Forecast performance*		
	StateVar	Bwidth	AvgLoss	GW stat	MCS
1*	time, RV	0.9995,0.31	0.315	−12.302	✓
2	RV	0.35	0.324	−11.726	×
3	time, VIX	0.9999, 0.27	0.339	−9.188	×
4	VIX	0.28	0.341	−9.050	×
5	time	0.995	0.370	−6.785	×
6	time, FFR	0.9975, 0.33	0.377	−6.489	×
7	time, 10Y-2Y	0.9975, 0.16	0.381	−4.286	×
8	10Y-2Y	0.11	0.396	−3.667	×
9	FFR	0.23	0.401	−4.348	×
10	–	–	0.414	★	×

Notes: This table presents measures of forecast performance over the out-of-sample period (January 2011 to June 2021) from GARCH(1,1) models estimated using either QML (non-local), or local QML. The rows are ordered by average OOS QLIKE loss, reported in the third-last column. The local method with the best performance in the validation sample (the second half of the estimation sample) is marked in the first column with ∗. The local estimators use the state variable(s) given in the second column and bandwidth parameter(s) from the third column, which are selected using the validation sample. All forecasting models are estimated using an expanding window of data. The penultimate column reports Giacomini-White t -statistics of each model relative to the benchmark non-local method (marked with ★), with negative t-statistics indicating lower average loss. The final column includes a check mark if a given method is included in the 95% model confidence set, and a cross otherwise.

In Fig. 3 we investigate whether the forecast gains from local estimation are sensitive to the choice of bandwidth. Similar to the lower panel of Fig. 1, we plot the average OOS loss from the local and non-local methods as a function of the bandwidth, as well as 95% GW confidence bands, indicating whether the two methods have significantly different average losses.[17] The optimal bandwidth from the validation sample is 0.35, while the *ex post* best bandwidth choice is 0.17. The difference in forecast improvement, however, is robust across a range of bandwidths, from approximately 0.12 to 0.46. As the bandwidth approaches zero the local forecast performance deteriorates, as estimation error swamps the forecast; as the bandwidth diverges the forecast improvements shrink to zero and the local and non-local forecasts coincide.

To quantify the *magnitude* of the forecast gains, in addition to their significance as gauged by GW tests, Supplemental Appendix SA.3 presents a method that draws on the trade-off between estimation error and goodness-of-fit that underlies our comparison of local and non-local models. We quantify the gains in forecast accuracy as a gain in effective sample size for estimating the benchmark model. Using that method we find that the improvement in forecast performance achieved by going from the benchmark GARCH model to the best local GARCH model is equivalent to the gain from using around 80 times more data for non-local estimation, indicating that the local GARCH model indeed produces forecasts that are substantially better than the benchmark GARCH model.

In Supplemental Appendix SA.4 we augment the benchmark GARCH model with an additional variable, making it a “GARCH-X” model. Given its usefulness as a state variable, we use VIX^2 as the “X” variable. (We use VIX^2 rather than VIX so that all regressors in the model are measures of variance.) Table S2 shows that 14 methods significantly (at the 0.05 level) beat the benchmark GARCH-X model, which ranks 15th out of the 20 competing methods. We find eight methods are included in the 95% MCS, and all of these methods are local, using VIX, RV, FFR and/or time as state variables. This confirms that the proposed local method improves the benchmark model even when an additional variable is included in that model, thereby altering the model, and also illustrates how to apply our method to an extension of a baseline model.

3.2. HAR volatility forecasts

We next consider a widely-used high frequency-based volatility forecasting model, the heterogeneous autoregressive (HAR) model of Corsi (2009). This model specifies one-period-ahead volatility to be a function of the one-day, one-week, and one-month lags of volatility:

$$RV_t = \beta_0 + \beta_d RV_{t-1} + \beta_w \frac{1}{5}\sum_{j=1}^{5} RV_{t-j} + \beta_m \frac{1}{22}\sum_{j=1}^{22} RV_{t-j} + e_t \tag{28}$$

By exploiting the information in high frequency data, this model has been widely found to out-perform the GARCH model based on daily data. We use five-minute realized volatility on the S&P 500 index over the period January 2000 to June 2021, and, as in the GARCH analysis in the previous section, we use 2000–2010 as the estimation sample (which is then further split into two to select the bandwidth parameters) and the remaining as the out-of-sample period. We also consider the same set of state variables: time, RV, VIX, FFR, 10Y-2Y, as well as bivariate state variables using time and each of the four stochastic state variables. We estimate the model using (local or non-local) QML.

[17] For ease of presentation, Fig. 3 is based on the local model estimated using RV as the state variable. This was the second-best model in the validation sample. The best model in the validation sample uses both time and RV as state variables, but visualizing the sensitivity to two bandwidths is more difficult. We find similar insensitivity in that case as in the case presented in Fig. 3.

Table 2
Out-of-sample forecast performance for HAR models.

Rank	*Method details*		*Forecast performance*		
	StateVar	Bwidth	AvgLoss	GW stat	MCS
1	time, VIX	0.9999, 0.35	0.238	−6.158	✓
2*	VIX	0.34	0.239	−5.802	×
3	time, 10Y-2Y	0.9999, 2.5	0.254	−2.393	×
4	time, RV	0.9999, 3.05	0.254	−2.445	×
5	time	0.9999	0.254	−2.867	×
6	time, FFR	0.9999, 2.5	0.254	−0.956	×
7	10Y-2Y	2.68	0.254	−1.436	×
8	RV	3.11	0.254	−0.474	×
9	–	–	0.254	★	×
10	FFR	2.51	0.255	1.875	×

Notes: This table presents measures of forecast performance over the out-of-sample period (January 2011 to June 2021) from HAR models estimated using either QML (non-local), or local QML. The rows are ordered by average OOS QLIKE loss, reported in the third-last column. The local method with the best performance in the validation sample (the second half of the estimation sample) is marked in the first column with ∗. The local estimators use the state variable(s) given in the second column and bandwidth parameter(s) from the third column, which are selected using the validation sample. All forecasting models are estimated using an expanding window of data. The penultimate column reports Giacomini-White t -statistics of each model relative to the benchmark non-local method (marked with ★), with negative t-statistics indicating lower average loss. The final column includes a check mark if a given method is included in the 95% model confidence set, and a cross otherwise.

Table 2 presents results on the out-of-sample forecast performance of the various estimation methods. The benchmark method ranks 9th out of the 10 estimators, and is significantly beaten, at the 0.05 level, by five local methods, based on RV, VIX, FFR and/or time. The local model using VIX is selected using the validation sample, and the GW statistic comparing this method to the benchmark is −5.8, strongly rejecting the benchmark in favor of the local estimator. Using the "equivalent sample size" method described in Supplemental Appendix SA.3, we find that the gain from using local estimation compared with non-local estimation is equivalent to the gain from using 45 times more data for non-local estimation, a substantial improvement. The 95% model confidence set contains just one estimator, the local method using time and VIX as state variables. These results reveal that even the more challenging HAR model can be improved by recognizing that it, too, is misspecified, and by tilting the parameters of the model to reflect the current environment as captured by the state variable.[18,19]

The theoretical analysis in Section 2.4 revealed that when a state variable that is only weakly related to the degree of misspecification in the model is considered, local estimation is likely to fare poorly compared with non-local estimation, as the deleterious effect of nonparametric estimation error will not be offset by improved fit. This appears to be the case in this application when using the Fed Funds Rate (FFR) as a state variable: when combined with time it performs better than the benchmark, though not significantly, and when used on its own the OOS average loss is greater than the benchmark, and has a GW t-statistic of 1.88, indicating significantly worse performance at the 0.10 level.

To illustrate how local and non-local estimation leads to different forecasts, Fig. 4 presents volatility forecasts over the last 18 months of the sample period obtained from the best local and non-local HAR models in Table 2. We see that for much of the period, volatility is low and the two methods yield very similar forecasts. The methods differ most markedly during the market turmoil in March 2020, where we observe that the local HAR produces forecasts that increase more quickly as market turbulence rose, and then decrease more quickly in the subsequent weeks.

3.3. VaR and ES forecasting

We now consider models for forecasting two key quantities in risk management: Value-at-Risk (VaR) and Expected Shortfall (ES). For a given probability level α, usually set at 5%, these two measures are defined as the α-quantile and the expected value conditional on being below the α-quantile, both conditional on information set $\mathcal{F}_{t-1}$:

$$Y_t|\mathcal{F}_{t-1} \sim F_t \tag{29}$$

$$\left[VaR_t, ES_t\right] \equiv \left[\ F_t^{-1}(\alpha)\ ,\ \mathbb{E}\left[Y_t|Y_t \le VaR_t, \mathcal{F}_{t-1}\right]\ \right] \tag{30}$$

While VaR is simply a quantile of the conditional distribution of the asset return under analysis, and thus estimation and forecasting of this measure can be done using the large literature on quantile forecasting (see Komunjer (2013), for a review), models for ES are relatively lacking. This is perhaps in part due to the fact that this risk measure is not "elicitable" (Gneiting, 2011), meaning that without strong assumptions there is no loss function that allows for its direct estimation. This hurdle was overcome by Fissler and

[18] Table S3 in the supplemental appendix presents results when the HAR-X model is taken as the baseline model. We find 15 methods significantly beat the benchmark, and the 95% MCS includes just two methods, both local versions of the HAR-X model using RV or time and RV as state variables.

[19] In Figure S1 in the Supplemental Appendix we show that the forecast gains from using a local HAR model are robust across a range of choices of the bandwidth parameter.

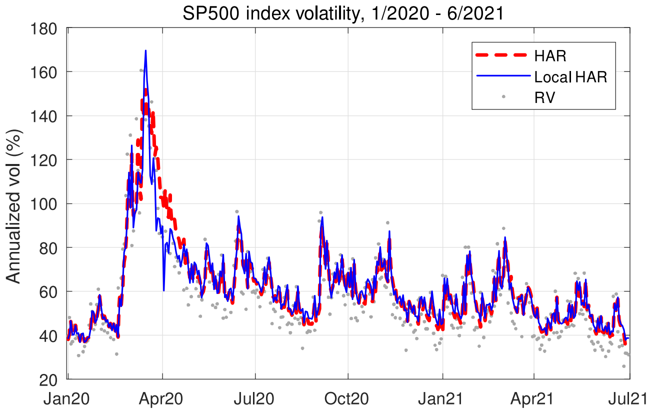

Fig. 4. This figure shows the predicted volatility from a HAR model estimated using local or non-local QML, along with realized volatility, over the last 18 months of the sample period.

Ziegel (2016), who proposed a class of loss functions that allows for the *joint* estimation of VaR and ES. We will focus on a leading member of this class, the "FZ0" loss function considered in Nolde and Ziegel (2017) and Patton et al. (2019):

$$L(y, v, e; \alpha) = -\frac{1}{\alpha e}\mathbf{1}\{y \leq v\}(v - y) + \frac{v}{e} + \log(-e) - 1 \tag{31}$$

With this loss function in hand, researchers can estimate models for VaR and ES directly (rather than indirectly via, for example, models for the entire predictive distribution) and competing forecasts of VaR and ES can be compared via their out-of-sample average FZ0 loss. Throughout, we consider a probability level, α, of 5%. This estimator can be interpreted as a QMLE, see Taylor (2019).

We take as the baseline model the zero-mean GARCH model, see Eq. (26). Using this model, forecasts for VaR and ES are obtained as:

$$\left[VaR_t, ES_t\right] = [a, b] \cdot \sigma_t \tag{32}$$

where $b < a < 0$ are the tail proportionality coefficients linking VaR and ES to volatility. If these parameters are estimated along with those of the GARCH model by minimizing the in-sample average FZ0 loss we obtain the "GARCH-FZ" model of Patton et al. (2019). We found that "localizing" these coefficients works poorly for forecasting, perhaps unsurprisingly as it combines nonparametrics and tail estimation, two data-intensive tasks.[20] Instead, we estimate $[a, b]$ using the standardized residuals based on the standard QML GARCH series, and only localize the GARCH model parameters. This leads to the GARCH-EDF model, Eq. (26) and:

$$\left[\hat{a}_t, \hat{b}_t\right] \equiv \left[\ \hat{F}_{\varepsilon,t}^{-1}(\alpha)\ ,\ \frac{1}{\alpha t}\sum_{s=1}^{t}\varepsilon_s \mathbf{1}\left\{\varepsilon_s \leq \widehat{VaR}_\varepsilon\right\}\ \right] \tag{33}$$

where $\varepsilon_t \equiv Y_t/\sigma_t$, $\hat{F}_{\varepsilon,t}^{-1}$ is the sample α-quantile of ε_t, and the GARCH process parameters are estimated by minimizing the FZ0 loss function. For the non-local estimation we obtain parameters by minimizing the in-sample average FZ0 loss function using an expanding window of data. For local M estimation, we follow the same method as in the previous sections: we consider a total of nine possible state variables, with bandwidth parameters optimized using the second half of the estimation sample.

Table 3 presents results on the out-of-sample forecast performance of the various estimation methods. The local method selected using the validation sample, which uses time and VIX as state variables, performs second-best in the OOS period. It significantly beats the benchmark, which is ranked 9th, though only at the 0.10 level with a GW statistic of −1.68. Using the "equivalent sample size" method described in Supplemental Appendix SA.3, we find that the gain from using local estimation compared with non-local estimation is equivalent to the gain from using just 1.5% more data for non-local estimation. Three other methods have significantly lower OOS than the benchmark, and all are local methods, using RV, VIX, or FFR as state variables. The MCS contains just a single method: local estimation using time and RV. Similar to the HAR application, we find that using the term spread (10Y-2Y) as a state variable worsens forecast performance, revealing that that state variable is only weakly related to the degree of misspecification GARCH-EDF model.

In Figure S1 in the Supplemental Appendix we examine the sensitivity of the forecast performance of the local GARCH-EDF model to the choice of the bandwidth parameter. In this figure we focus on the case that the state variable is time, and the bandwidths range from 0.98 to 0.9999. When the bandwidth is chosen "too small," in this case meaning in the lower three-quarters of the range considered, the local model performance deteriorates, and for values in the lowest quarter of this range the difference from the non-local forecasts are statistically significant. However, the model's performance is stable across a range of bandwidths near the validation sample-optimal bandwidth.

[20] Table S4 in the supplemental appendix is analogous to Table 3, discussed below, using the GARCH-FZ as the benchmark model. There we see that some local methods significantly beat the non-local benchmark, but overall the performance is worse, and for this reason we focus on GARCH-EDF as the baseline model.

Table 3
Out-of-sample forecast performance for VaR-ES models.

Rank	*Method details*		*Forecast performance*		
	StateVar	Bwidth	AvgLoss	GW stat	MCS
1	time, RV	0.9975. 2.57	−3.868	−1.987	✓
2*	time, VIX	0.9975, 1.68	−3.864	−1.678	×
3	time	0.9975	−3.860	−1.329	×
4	time, FFR	0.9975, 3.09	−3.860	−1.289	×
5	time, 10Y-2Y	0.9975, 2.52	−3.859	−1.206	×
6	RV	2.06	−3.852	−5.682	×
7	VIX	2.01	−3.849	−3.815	×
8	FFR	2.97	−3.845	−2.253	×
9	–	–	−3.843	★	×
10	10Y-2Y	3.18	−3.843	0.177	×

Notes: This table presents measures of forecast performance over the out-of-sample period (January 2011 to June 2021) from GARCH(1,1) models estimated either M estimation or local M estimation and the FZ0 loss function in Eq. (31). The rows are ordered by average OOS FZ0 loss, reported in the third-last column. For a given model, the local method with the best performance in the validation sample (the second half of the estimation sample) is marked in the first column with $*$. The local estimators use the state variable(s) given in the second column and bandwidth parameter(s) from the third column, which are selected using the validation sample. All forecasting models are estimated using an expanding window of data. The penultimate column reports Giacomini-White t-statistics of each model relative to the benchmark non-local method (marked with ★), with negative t-statistics indicating lower average loss. The final column includes a check mark if a given method is included in the 95% model confidence set, and a cross otherwise.

3.4. *Yield curve forecasting*

In our final empirical application we consider the popular "dynamic Nelson–Siegel" model for predicting the term structure of bond yields proposed by Diebold and Li (2006). Denoting $y_t(\tau)$ as the yield on a bond with maturity τ at time t, this model starts from the Nelson and Siegel (1987) model for a term structure of yields:

$$y_t(\tau) = \beta_{1,t} + \beta_{2,t}\left(\frac{1-\exp\left\{-\lambda_t\tau\right\}}{\lambda_t\tau}\right) + \beta_{3,t}\left(\frac{1-\exp\left\{-\lambda_t\tau\right\}}{\lambda_t\tau} - \exp\left\{-\lambda_t\tau\right\}\right) + e_t \tag{34}$$

This specification has four free parameters: the betas affect the level, slope and curvature of the yield curve, while λ_t determines (among other things) the maturity at which the curvature factor has a turning point. These parameters can be estimated jointly, period-by-period, using nonlinear least squares, or if λ_t is fixed at some pre-determined value the remaining parameters can be obtained analytically using OLS. We follow Diebold and Li (2006) and set $\lambda_t = 0.0609\ \forall\ t$ so that the curvature term peaks at 30 months and the model can be estimated by OLS.

Moving beyond describing yield curves to predicting them, Diebold and Li (2006) proposed modeling the observed sequences of $\left\{\beta_{i,t}\right\}_{t=1}^{T}$, for $i = 1, 2, 3$, as AR(1) processes:

$$\beta_{\,i,t+1} = \phi_{0i} + \phi_{1i}\beta_{i,t} + e_{i,t+1} \tag{35}$$

That is, on each day in the estimation window the vector $\left[\beta_{1,t}, \beta_{2,t}, \beta_{3,t}\right]$ is obtained from the cross-section of yields, and then from the time series of these parameters the predicted value of the vector for the next period is obtained by estimating an AR(1) model via OLS. Inserting those forecasts into the Nelson–Siegel functional form then provides a forecast for the next-period yield curve, and combined the Eqs. (34) and (35) comprise the "dynamic Nelson–Siegel" (DNS) model.

We consider local versions of the DNS model, where the three AR(1) models are estimated via local OLS based on one of the nine state variables used in the previous analyses. Local OLS estimation of this model simplifies to weighted OLS (see, e.g., Cleveland and Devlin (1988) and Fan et al. (1998)), with the weights coming directly from the state variable and the kernel, and, as for OLS, the estimated local parameters are available in closed form, see Supplemental Appendix SA.1 for more details. We use the same state variable (and same bandwidth value) for all three AR(1) models, although that could be relaxed.[21] We additionally consider the usual, non-local, DNS model, estimated on an expanding window of data.

We use daily data over the period January 2000 to June 2021, and we consider bonds with maturities of three and six months, and one to ten years, a total of twelve maturities.[22] We summarize the predictive performance of this model by summing the squared OOS forecast errors across maturities.

Table 4 presents the results for two forecast horizons, one day and twenty days. The results in Panel A, for the one-day horizon, show that the benchmark method is ranked last out of the ten methods considered, and is statistically significantly beaten by all nine competitors. However, as the third-last column in Table 4 reveals, the RMSEs from the competing models are identical up to

[21] We choose the bandwidth to minimize the sum of the MSEs across the three AR(1) models, however it is possible to consider different state variables, with different bandwidths, for each of the three AR(1) models for the betas. We have not considered this extension.

[22] We obtain one- to ten-year yields from https://www.federalreserve.gov/data/nominal-yield-curve.htm and data on three- and six-month yields, as well as the FFR and 10Y-2Y from the St. Louis Fed "FRED" database.

Table 4
Out-of-sample forecast performance for yield curve models.

Rank	*Method details*		*Forecast performance*		
	StateVar	Bwidth	AvgLoss	GW stat	MCS
Panel A: One-day forecast horizon					
1	time, RV	0.9999, 1.2	0.158	−9.471	✓
2	time, FFR	0.9999, 0.74	0.158	−5.736	✓
3	RV	1.19	0.158	−7.522	×
4	FFR	0.74	0.158	−4.698	×
5	time, 10Y-2Y	0.9999, 0.49	0.158	−6.023	×
6	time, VIX	0.9999, 1.53	0.158	−5.748	×
7*	10Y-2Y	0.49	0.158	−3.544	×
8	time	0.9999	0.158	−16.634	×
9	VIX	1.5	0.158	−2.365	×
10	–	–	0.158	★	×
Panel B: Twenty-day forecast horizon					
1	time, FFR	0.9999, 3.2	0.244	−6.966	✓
2	time, 10Y-2Y	0.9999, 0.76	0.244	−1.914	✓
3	time	0.9999	0.245	−7.294	×
4	time, VIX	0.9999, 1.95	0.245	−3.352	×
5	FFR	3.2	0.245	−6.146	×
6*	10Y-2Y	0.76	0.246	−0.426	×
7	time, RV	0.9999, 2.08	0.246	−0.496	×
8	–	–	0.246	★	×
9	VIX	1.98	0.246	0.648	×
10	RV	2.11	0.247	4.477	×

Notes: This table presents measures of one- and twenty-day-ahead forecast performance over the out-of-sample period (January 2011 to June 2021) from dynamic Nelson–Siegel models estimated using either OLS or local OLS. The rows in each panel are ordered by average OOS RMSE, multiplied by 100, reported in the third-last column. The local method with the best performance in the validation sample is marked in the first column with ∗. The local estimators use the state variable(s) given in the second column and bandwidth parameter(s) from the third column, which are selected using the validation sample. All forecasting models are estimated using an expanding window of data. The penultimate column reports Giacomini-White t-statistics of each model relative to the benchmark non-local method (marked with ★), with negative t-statistics indicating lower average loss. The final column includes a check mark if a given method is included in the 95% model confidence set, and a cross otherwise.

the first three decimal places. The improvement in RMSE of the best local method relative to the benchmark is just 0.1%. Moreover, the "equivalent sample size" measure of the improvement from using local estimation is a gain in sample size of just 16%. Both of these suggest that despite the statistical significance, the economic improvement from local estimation is limited. This negative result connects to the theoretical analysis in Section 2.4, in that the best non-local method in this application has an R^2 of 0.964, leaving very little room for improvement by a competing method.

In Panel B of Table 4 we present results for the 20-day horizon, and for this more challenging forecasting problem we see that local estimation leads to improved OOS performance. The benchmark method ranks eighth out of the ten estimators, and it is significantly beaten by four local methods: those based on time alone, and time combined with FFR, 10Y-2Y or VIX. The two methods in the MCS use time and FFR or 10Y-2Y as state variables. Using the "equivalent sample size" method described in Supplemental Appendix SA.3, we find that the gain from using local estimation compared with non-local estimation is equivalent to the gain from using 4 times more data for non-local estimation, a substantial improvement.[23] Panel B of Table 4 provides another example of a poor state variable leading to worse out-of-sample forecast performance: local estimation using either RV or VIX leads to higher RMSE, and in the case of RV the difference is strongly significant, with a GW statistic of over 4.

Combined, the results from the yield curve forecasting application highlight the upsides and the downsides of local estimation. When the baseline model is very good, as it is for the one-day forecast horizon, there is little scope for an alternative estimation method to offer any gains. However for more difficult forecasting problems, alternative estimation methods like the local methods proposed here offer the possibility of improved forecasts, so long as the state variable is informative about the benchmark model's misspecification.

3.5. Conditional comparisons of forecast performance

In all of the above analyses we focused on the *average* out-of-sample (OOS) performance of local and non-local methods for estimating a forecasting model. However, if the forecast user has an idea for a state variable that may be useful for tilting the estimated model parameters, this variable may also be useful for predicting which method is likely to outperform in the next period. We investigate this idea in three ways: via linear regression, nonparametric regression, and a test of uniform predictive performance.

[23] In Figure S2 in the Supplemental Appendix we show that the results for both yield curve applications are robust across a range of choices of the bandwidth parameter.

Table 5
Conditional comparisons of forecasting models.

	GARCH	HAR	VaR-ES	Yield curve	
				h = 1	h = 20
Intercept	−0.099	−0.016	−0.020	−0.052	−0.373
(std. err.)	(0.008)	(0.003)	(0.012)	(0.015)	(0.875)
[t-stat]	[−12.302]	[−5.802]	[−1.678]	[−3.544]	[−0.426]
Slope	0.090	−0.014	−0.000	0.087	−2.901
(std. err.)	(0.009)	(0.010)	(0.056)	(0.076)	(7.591)
[t-stat]	[10.297]	[−1.361]	[−0.003]	[1.155]	[−0.382]

Notes: This table presents the estimated parameters and standard errors from a linear regression of out-of-sample loss differences on a constant and the lagged state variable, across the five applications considered in this paper. The methods compared in each column are the local method with the best performance in the validation sample (marked with $*$ in each of Tables 1 to 4) and the non-local method using the full estimation sample. The state variable used for the comparison is the same one that appears in the local method: RV for the GARCH application, VIX for the HAR and VaR-ES application, and 10Y-2Y for the yield curve (h = 1 and h = 20) applications.

In each case we compare the local method with the best performance in the validation sample (these are marked with $*$ in each of Tables 1 to 4) to the benchmark non-local method. The state variable used is the same as that in the local method: RV for the GARCH application, and VIX for the HAR and VaR-ES applications, and the slope of the term structure (10Y-2Y) for the yield curve (h = 1 and h = 20) applications.

Table 5 presents the results of a simple linear regression of OOS loss differences on a constant and the lagged state variable, as proposed in Giacomini and White (2006). We de-mean the state variable so that the intercept of this regression corresponds to the difference in average OOS loss, and the t-statistics associated with the intercept are exactly the GW statistics for the unconditional comparisons in Tables 1 to 4. The t -statistics on the slope coefficient reveal whether the state variable can (linearly) predict future differences in realized losses. In the GARCH application the slope coefficient (on RV) is positive and significant, indicating that the local method does relatively worse when volatility is high. In the HAR and VaR-ES applications the slope (on VIX) is negative, but not significant, indicating that the local method does relatively better when volatility is high. The slope coefficient switches signs in the two yield curve applications, but is not significant in either.

To gain a more nuanced understanding of the relationship between OOS loss differences and the state variable, Figs. 5 and 6 present a simple nonparametric kernel smooth of this relationship, along with pointwise 95% confidence intervals.[24] These plots allow us to see if the loss difference particularly positive or negative in some part of the support of the state variable. In the upper panel of Fig. 5 we see that local QML strongly outperforms non-local QML for GARCH models when volatility is relatively low. When annualized RV is above about 15% the difference in performance is approximately zero, and the confidence interval includes zero for all values of RV above 20%. Similar results hold for the VaR-ES comparison.

For the HAR application, presented in the middle panel of Fig. 5 , we see that the predicted OOS loss difference is almost constant in the state variable. The loss difference is always negative, revealing that local QML outperforms non-local QML regardless of the level of volatility, though the significance of the difference drops as volatility rises above about 30%.

In the upper panel of Fig. 6, we see that local OLS significantly outperforms non-local OLS when the term structure is relatively flat (when the difference between 10-year and 2-year government bonds is near zero). When the term structure steepens to greater than about 1%, the difference in performance is not significant. In the lower panel of Fig. 6 we observe the reason for the insignificant slope coefficient in the linear analysis presented in Table 5: the relationship is U-shaped. When the term structure is relatively flat or relatively steep, non-local OLS weakly dominates local OLS, while for intermediate values of the slope (between 0.25% and 2.5%) local OLS significantly outperforms non-local OLS.[25]

Finally, we use the recently proposed "conditional superior predictive ability" (CSPA) test of Li et al. (2022) to test whether the non-local method has weakly lower expected loss across the entire support of the state variable:

$$H_0 : \mathbb{E}\left[L\left(Y_{t+1}, g_t\left(\tilde{\theta}_{h,t}\left(S_t\right)\right)\right) - L\left(Y_{t+1}, g_t(\hat{\theta}_t)\right) \middle| S_t = s\right] \geq 0 \;\forall\; s \in \mathrm{Int}\left(\mathcal{S}\right) \tag{36}$$

as well the hypothesis where the inequality in Eq. (36) is reversed. In the GARCH application, we reject the first null (p-value less than 0.01) and conclude that non-local QML does not weakly dominate local QML uniformly, which is unsurprising given the estimated average loss presented in the upper panel of Fig. 5. We fail to reject the reverse hypothesis (p-value of 0.99), meaning that local QML may indeed dominate non-local QML, and combined these results indicate that local QML is strongly preferred to non-local QML. We find the same outcomes for the HAR and both yield curve ($h = 1$ and $h = 20$) applications: local estimation is strongly preferred to non-local estimation. In contrast, in the VaR-ES application we fail to reject either null at the 0.05 level, despite local estimation outperforming non-local estimation unconditionally, and outperforming pointwise for low values of VIX as in Fig. 5. This outcome may be due to a relative lack of power in this application, which is focused on the 5% tail of the distribution of returns.

[24] The estimate and confidence intervals are computed using Theorem 2.2 of Li and Racine (2007).

[25] It is possible to construct a "hybrid" forecast based on the local and non-local methods by switching between them according to which method is predicted to have lower loss in the subsequent period, see Giacomini and White (2006) and Zhu and Timmermann (2022) for example. We do not pursue this extension here.

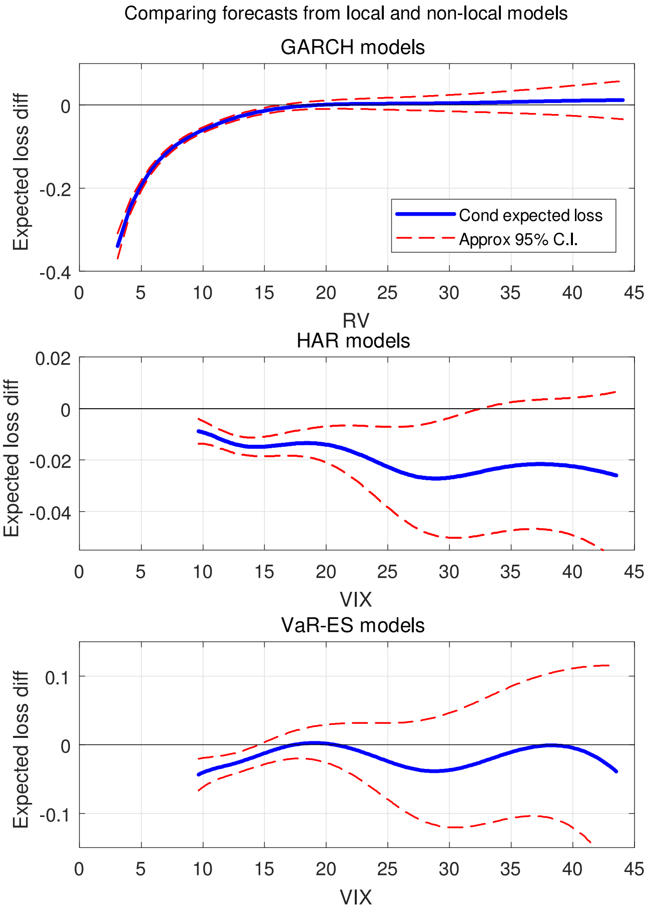

Fig. 5. This figure presents estimates of the expected out-of-sample loss differences from models estimated via local or non-local methods, conditional on realized volatility (top panel) or VIX (lower two panels). Positive loss differences indicate the non-local method is preferred.

4. Conclusion

This paper proposes an estimation method to improve the forecasts produced by a misspecified forecasting model, without altering the form of the underlying model. In many decision-making environments, the statistical model is “hardwired,” at least in the short term, and substituting it for a new and improved model is not possible. This may be because changing the model requires regulatory approval, or approval from a high-level committee, or because the time taken to embed a new model in the decision-making process is long relative to the competitive environment. We overcome this hurdle by maintaining the functional form of the baseline model and improving its fit by upweighting past observations that look more similar to the forecast date, and downweighting observations that are more dissimilar, drawing on methods like local OLS estimation and local MLE, see Tibshirani and Hastie (1987), Cleveland and Devlin (1988) and Fan et al. (1998), as well as older methods like exponential smoothing, see Brown (1956) and Muth (1960).

We theoretically compare out-of-sample forecasts from the proposed estimation method with those from the baseline model and observe a familiar bias–variance trade-off. Interestingly, the bias–variance trade-off for the proposed method goes in the opposite direction to the usual one for out-of-sample forecasting: the proposed estimation method (generally) adds variance to the forecast, in the hope of reducing the bias from using the misspecified baseline model. Our theoretical analysis sheds light on the conditions that are likely to be favorable for the local estimation method proposed here. Specifically, the baseline model cannot be “too good” and the forecaster’s state variable summarizing the environment at the forecast date cannot be “too bad.”

We apply the proposed method to four economic forecasting problems. The first two applications consider volatility forecasting, using daily data and the famous GARCH model of Bollerslev (1986) or high frequency data and the popular HAR model of Corsi (2009). The third application is to risk management, and focuses on joint forecasts of Value-at-Risk and Expected Shortfall. The fourth application is to yield curve forecasts, made using the “dynamic Nelson–Siegel” model proposed by Diebold and Li (2006). We find that our proposed method provides statistically significant improvements over the baseline methods in almost all cases.

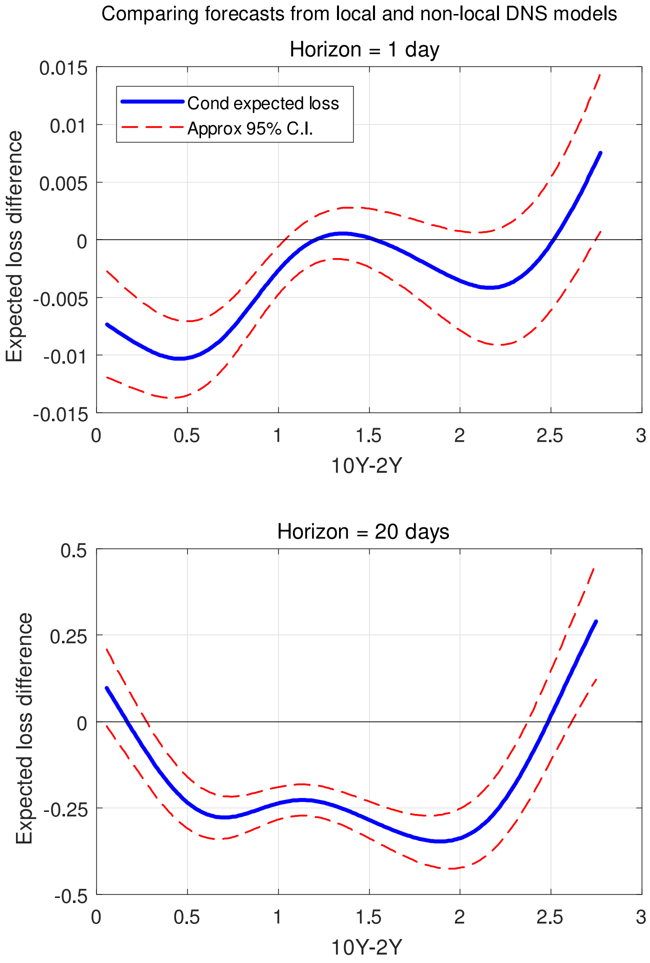

Fig. 6. This figure presents estimates of the expected out-of-sample loss difference of a dynamic Nelson–Siegel (DNS) model estimated via local OLS or non-local OLS, both conditional on the difference between 10-year and 2-year government bond yields (10Y-2Y). Positive loss differences indicate the non-local method is preferred.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2024.105767.

References

Andersen, T.G., Bollerslev, T., Diebold, F.X., 2007. Roughing it up: Disentangling continuous and jump components in measuring, modeling and forecasting asset return volatility. Rev. Econ. Stat. 89 (4), 701–720.

Ang, A., Bekaert, G., Wei, M., 2007. Do macro variables, asset markets, or surveys forecast inflation better? J. Monetary Econ. 54 (4), 1163–1212.

Ang, A., Kristensen, D., 2012. Testing conditional factor models. J. Financ. Econ. 106, 132–156.

Beare, B.K., 2010. Copulas and temporal dependence. Econometrica 78, 395–410.

Blasques, F., Koopman, S.J., Mallee, M., Zhang, Z., 2016. Weighted maximum likelihood for dynamic factor analysis and forecasting with mixed frequency data. J. Econometrics 193, 405–417.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econometrics 31, 307–327.

Brown, R.G., 1956. Exponential Smoothing for Predicting Demand. Arthur D. Little Inc., Cambridge, Massachusetts.

Chen, X., Fan, Y., 2006. Estimation of copula-based semiparametric time series models. J. Econometrics 130, 307–335.

Christoffersen, P., Jacobs, K., Ornthanalai, C., Wang, Y., 2008. Option valuation with long-run and short-run volatility components. J. Financ. Econ. 90, 272–297.

Cleveland, W.S., Devlin, S.J., 1988. Locally weighted regression: An approach to regression analysis by local fitting. J. Amer. Statist. Assoc. 83, 596–610.

Corsi, F., 2009. A simple approximate long-memory model of realized volatility. J. Financ. Econom. 7 (2), 174–196.

Dendramis, Y., Kapetanios, G., Marcellino, M., 2020. A similarity-based approach for macroeconomic forecasting. J. Roy. Statist. Soc. Ser. A 183 (3), 801–827.

Diebold, F.X., Li, C., 2006. Forecasting the term structure of government bond yields. J. Econometrics 130, 337–364.

Engle, R.F., Lee, G.G.J., 1999. A permanent and transitory component model of stock return volatility. In: Engle, R.F., White, H. (Eds.), Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W. J. Granger. Oxford University Press, pp. 475–497.

Fan, J., Farmen, M., Gijbels, I., 1998. Local maximum likelihood estimation and inference. J. R. Stat. Soc., Ser. B 60 (3), 591–608.

Fan, J., Wu, Y., Feng, Y., 2009. Local quasi-likelihood with a parametric guide. Ann. Stat. 37 (6B), 4153–4183.
Fan, J., Yao, Q., 2003. Nonlinear Time Series: Nonparametric and Parametric Methods. Springer, New York.
Faust, J., Wright, J.H., 2009. Comparing greenbook and reduced form forecasts using a large realtime dataset. J. Bus. Econom. Statist. 27 (4), 468–479.
Fissler, T., Ziegel, J.F., 2016. Higher order elicitability and Osband's principle. Ann. Statist. 44 (4), 1680–1707.
Giacomini, R., Ragusa, G., 2014. Theory-coherent forecasting. J. Econometrics 182, 145–155.
Giacomini, R., Rossi, B., 2010. Forecast comparisons in unstable environments. J. Appl. Econometrics 25 (4), 595–620.
Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74, 1545–1578.
Gneiting, T., 2011. Making and evaluating point forecasts. J. Amer. Statist. Assoc. 106, 746–762.
Granger, C.W.J., 1969. Prediction with a generalized cost of error function. OR 20 (2), 199–207.
Hansen, P.R., Dumitrescu, E.-I., 2022. How should parameter estimation be tailored to the objective? J. Econometrics 230, 535–558.
Hansen, P.R., Lunde, A., 2005. A forecast comparison of volatility models: Does anything beat a GARCH (1, 1)? J. Appl. Econometrics 20 (7), 873–889.
Hansen, P.R., Lunde, A., Nason, J.M., 2011. The model confidence set. Econometrica 79 (2), 453–497.
Härdle, W., Tsybakov, A., 1997. Local polynomial estimators of the volatility function in nonparametric autoregression. J. Econometrics 81, 223–242.
Härdle, W., Tsybakov, A., Yang, L., 1998. Nonparametric vector autoregression. J. Statist. Plann. Inference 68 (2), 221–245.
Hu, F., 1997. The asymptotic properties of the maximum-relevance weighted likelihood estimators. Canad. J. Statist. 25 (1), 45–59.
Inoue, A., Jin, L., Pelletier, D., 2020. Local-linear estimation of time-varying-parameter GARCH models and associated risk measures. J. Financ. Econom. 19 (1), 202–234.
Inoue, A., Jin, L., Rossi, B., 2017. Rolling window selection for out-of-sample forecasting with time-varying parameters. J. Econometrics 196, 55–67.
Komunjer, I., 2013. Quantile prediction. In: Elliott, G., Timmermann, A. (Eds.), In: Handbook of Economic Forecasting, vol. 2, Elsevier, Oxford, pp. 961–994.
Kristensen, D., Mele, A., 2011. Adding and subtracting Black–Scholes: A new approach to approximating derivative prices in continuous-time models. J. Financ. Econ. 102, 390–415.
Li, J., Liao, Z., Quaedvlieg, R., 2022. Conditional superior predictive ability. Rev. Econ. Stud. 89 (2), 843–875.
Li, Q., Racine, J.S., 2007. Nonparametric Econometrics. Princeton University Press, Princeton.
Manganelli, S., 2009. Forecasting with judgment. J. Bus. Econom. Statist. 27 (4), 553–563.
Muth, J.F., 1960. Optimal properties of exponentially weighted forecasts. J. Amer. Statist. Assoc. 55 (290), 299–306.
Nelson, C.R., Siegel, A.F., 1987. Parsimonious modeling of yield curve. J. Bus. 60, 473–489.
Newey, W.K., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, Vol. 4. North-Holland, Amsterdam.
Nolde, N., Ziegel, J.F., 2017. Elicitability and backtesting: Perspectives for banking regulation. Ann. Appl. Stat. 11 (4), 1833–1874.
Patton, A.J., 2020. Comparing possibly misspecified forecasts. J. Bus. Econom. Statist. 38 (4), 796–809.
Patton, A.J., Ziegel, J.F., Chen, R., 2019. Dynamic semiparametric models for expected shortfall (and value-at-risk). J. Econometrics 211 (2), 388–413.
Pesaran, M.H., Pick, A., Pranovich, M., 2013. Optimal forecasts in the presence of structural breaks. J. Econometrics 177, 134–152.
Pettenuzzo, D., Timmermann, A., Valkanov, R., 2014. Forecasting stock returns under economic constraints. J. Financ. Econ. 114, 517–553.
Richter, S., Smetanina, E., 2020. Forecast evaluation and selection in unstable environments. working paper, Chicago Booth.
Taylor, J.W., 2019. Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. J. Bus. Econom. Statist. 37 (1), 121–133.
Tibshirani, R., Hastie, T., 1987. Local likelihood estimation. J. Am. Stat. Assoc. 82 (398), 559–567.
Weiss, A.A., 1996. Estimating time series models using the relevant cost function. J. Appl. Econometrics 11 (5), 539–560.
White, H., 1994. Estimation, Inference and Specification Analysis. Cambridge University Press, Cambridge, U.K..
White, H., 2001. Asymptotic Theory for Econometricians. Academic Press, San Diego, California.
Zhu, Y., Timmermann, A., 2022. Conditional rotation between forecasting models. J. Econometrics 231, 329–347.
Zumbach, G., 2006. The RiskMetrics 2006 methodology. working paper, RiskMetrics Group, Geneva, Switzerland.