Finance and Economics Discussion Series

Federal Reserve Board, Washington, D.C.
ISSN 1936-2854 (Print)
ISSN 2767-3898 (Online)

Better the Devil You Know: Improved Forecasts from Imperfect Models

Dong Hwan Oh and Andrew J. Patton

2021-071

Please cite this paper as:
Oh, Dong Hwan, and Andrew J. Patton (2021). "Better the Devil You Know: Improved Forecasts from Imperfect Models," Finance and Economics Discussion Series 2021-071. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2021.071.

Better the Devil You Know: Improved Forecasts from Imperfect Models*

Dong Hwan Oh
Federal Reserve Board

Andrew J. Patton
Duke University

First draft: August 2021. This draft: October 2021.

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 distinct empirical analyses including volatility forecasting, risk management, and yield curve forecasting.

Keywords: model misspecification, local maximum likelihood, volatility forecasting, value-at-risk and expected shortfall forecasting, yield curve forecasting.

J.E.L. codes: C53, C51, C58, C14.

*We thank Tim Bollerslev, Ana Galvao, Mike McCracken, Rogier Quaedvlieg, Allan Timmermann (discussant) and participants in the SoFiE Seminar Series. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Federal Reserve Board. Email: donghwan.oh@frb.gov, andrew.patton@duke.edu.

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.

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

[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.

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. Gia-

[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.

comini 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.

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:[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 $t = 0, 1, ..., T$, 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, naturally, $\mathcal{F}_t$-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, and $\text{Int}(\mathcal{S})$ is the interior of the set $\mathcal{S}$. Under a variety of regularity conditions, see e.g. Fan *et al.* (2009), the limit of this estimator is:

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

[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] Matching the estimation and evaluation loss functions is intuitive and can lead to improved forecasts, see Granger (1969) and Weiss (1996) for example, however in some applications there may be gains from using an alternative loss function for estimation, see Hansen and Dumitrescu (2021).

With the bandwidth shrinking at an appropriate rate, which differs depending on assumptions about smoothness and temporal dependence, 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.[6]

2.2 The special case of correct specification

Consider the special case that the baseline model is correctly specified and point identified for the target functional. This implies

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

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

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

which implies

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

However, 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

$$\mathbb{E}\left[L\left(Y_{t+1}, g_t^\dagger\right)|\mathcal{F}_t\right] = \mathbb{E}\left[L\left(Y_{t+1}, g_t(\hat{\theta}^*)\right)\middle|\mathcal{F}_t\right] \leq \mathbb{E}\left[L\left(Y_{t+1}, g_t(\theta)\right)|\mathcal{F}_t\right] \ \text{ a.s. } \forall\ t,\ \forall\ \theta \in \Theta \tag{11}$$

[6] 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.

Since $S_t \in \mathcal{F}_t$, we can combine Equation (10) with the law of iterated expectations (LIE) to infer

$$\mathbb{E}\left[L\left(Y_{t+1}, g_t(\hat{\theta}^*)\right)\middle| S_t\right] = \mathbb{E}\left[L\left(Y_{t+1}, g_t\left(\tilde{\theta}^*\left(S_t\right)\right)\right)\middle| S_t\right] \quad \text{a.s. } \forall\ t \tag{12}$$

and thus, by the point-identification assumption, that $\tilde{\theta}^*\left(S_t\right) = \hat{\theta}^*$. Noting that this must be true (a.s.) for all t, this implies that $\tilde{\theta}^*\left(s\right)$ 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 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.

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 simply shows that the OOS average loss from the local estimator will weakly dominate that from the usual estimator in population.[7,8] The gains accrue as 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 $\tilde{\theta}^*\left(s\right) = \hat{\theta}^*\ \forall\ s$ 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

[7] Note that this is true even though OOS performance is computed using *non*-weighted losses, that is, the kernel function used in the local estimator does not appear.

[8] 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. (2021), 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, but our focus in this section is on overall (unconditional) performance in the OOS period.

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 $\tilde{\theta}_{h,T}(S_T)$, centered on the limiting parameter $\tilde{\theta}^*(S_T)$. For ease of presentation we assume that $\dim(\theta) = 1$, which can easily be relaxed, and we assume that the loss function is differentiable.[9]

$$
\begin{aligned}
& L\left(Y_{T+1}, g_T\left(\tilde{\theta}_{h,T}(S_T)\right)\right) && (15)\\
\approx\ & L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*(S_T)\right)\right) + \frac{\partial L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*(S_T)\right)\right)}{\partial\theta}\left(\tilde{\theta}_{h,T}(S_T) - \tilde{\theta}^*(S_T)\right) && (16)\\
& + \frac{1}{2}\frac{\partial^2 L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*(S_T)\right)\right)}{\partial\theta^2}\left(\tilde{\theta}_{h,T}(S_T) - \tilde{\theta}^*(S_T)\right)^2
\end{aligned}
$$

Taking conditional expectations we then find

$$
\begin{aligned}
\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}_{h,T}(S_T)\right)\right)\middle|\mathcal{F}_T\right] \approx\ & \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*(S_T)\right)\right)\middle|\mathcal{F}_T\right] && (17)\\
& + \tilde{H}_T^*(S_T)\left(\tilde{\theta}_{h,T}(S_T) - \tilde{\theta}^*(S_T)\right)^2 \\
\text{where}\quad \tilde{H}_T^*(S_T) \equiv\ & \mathbb{E}\left[\frac{1}{2}\frac{\partial^2 L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*(S_T)\right)\right)}{\partial\theta^2}\middle|\mathcal{F}_T\right] && (18)
\end{aligned}
$$

The first-order term in the expansion drops out as $\mathbb{E}\left[\partial L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*(S_T)\right)\right)/\partial\theta\middle|\mathcal{F}_T\right] = 0$ a.s. by the definition of $\tilde{\theta}^*(S_T)$. $\tilde{H}_T^*(S_T)$ is a Hessian-like term and is positive definite in standard estimation problems. Then, taking unconditional expectations we obtain

$$
\begin{aligned}
\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}_{h,T}(S_T)\right)\right)\right] \approx\ & \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*(S_T)\right)\right)\right] && (19)\\
& + \mathbb{E}[\,\tilde{H}_T^*(S_T)\left(\tilde{\theta}_{h,T}(S_T) - \tilde{\theta}^*(S_T)\right)^2]
\end{aligned}
$$

The first term on the right-hand side is the OOS average loss for the local estimator evaluated at the population parameter, and the second term is positive and of the order $\mathcal{O}_p\left(T^{-1+2\gamma}\right)$.

[9] If the target variable is continuously distributed, non-differentiable loss functions like the Fissler-Ziegel loss used in Value-at-Risk and Expected Shortfall estimation, can be accommodated by approximating the expected loss.

Next we consider the usual, non-local, estimator using similar steps, and obtain:

$$\begin{aligned}\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\hat{\theta}_T\right)\right)\right] &\approx \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\hat{\theta}^*\right)\right)\right] + \mathbb{E}[\hat{H}_T^*\left(\hat{\theta}_T - \hat{\theta}^*\right)^2] \\ \text{where } \hat{H}_T^* &\equiv \mathbb{E}\left[\frac{1}{2}\frac{\partial^2 L\left(Y_{T+1}, g_T\left(\hat{\theta}^*\right)\right)}{\partial\theta^2}\middle| \mathcal{F}_T\right]\end{aligned}$$

The expected loss using the estimated parameter is again equal to the average loss based on the infeasible population parameter, and a positive term related to estimation error. In this case, the estimation error term is of order $\mathcal{O}_p\left(T^{-1}\right)$.

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

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

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, which is of the order $\mathcal{O}_p\left(T^{-1+2\gamma}\right)$. Since this term is positive, it increases the expected loss 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 $\tilde{\theta}^*\left(s\right) = \hat{\theta}^*\ \forall\ s$, 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.,

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

then the local estimation's FOC is satisfied when $\tilde{\theta}^*(S_t) = \hat{\theta}^*$, 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 $\tilde{\theta}^*(S_t) = \hat{\theta}^*$ in this case, there is 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

[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.

copula linking adjacent realizations:

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

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 Figure 1 we see that it is increasing and concave. The upper panel of Figure 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 then 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 Figure 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 Figure 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.

[INSERT FIGURE 1 ABOUT HERE]

[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 Figure 1.

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:

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

We consider values for the bandwidth, h, in the range $0.01\sigma_S$ to $3\sigma_S$, 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:

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

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 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 are re-estimated daily throughout the OOS period using a rolling window of data, while bandwidth parameters are kept fixed at their optimized value from the validation sample.

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,[13] and 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.

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-

[12] For the bandwidth h we use a coarse grid of width 0.1 from $0.1\sigma_S$ to $3\sigma_S$ 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] This data is taken from the Oxford-Man Realised Library.

variance trade-off relative to the benchmark method, in the spirit of the analysis in Section 2. We use Giacomini-White (2006) (GW) tests to compare each method to the non-local method using the full sample for estimation, and we estimate the set of best methods using the model confidence set (MCS) of Hansen *et al.* (2011).[14] 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.[15] 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.

[14] 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.

[15] There are many papers that have built on the original GARCH model, we do not attempt to conduct a horserace of volatility models here. Rather we illustrate how our method improves upon on 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.

As described above, we consider a total of nine possible state variables for local estimation. For non-local estimation we consider estimation windows of length 250, 500, 1000 and the full estimation period (2737 observations). By considering both long and short estimation windows for the baseline model, we can see whether the proposed local method out-performs a well-known way of "localizing" estimation; namely, using a shorter estimation window.

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 ($\checkmark$ or $\times$) 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 the full estimation window, is ranked *last* in this set of estimation methods. Every local method aside from those using the two yield curve state variables (FFR and 10Y-2Y) has significantly lower OOS loss than the benchmark method, according to the GW test. 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 non-local method with the local method selected using the validation sample we obtain a GW t-statistic of -10.32, strong evidence that the local method out-performs the non-local benchmark.[16] 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 VIX and time as the state variables. This small MCS indicates a high degree of precision in identifying the best-performing method.

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

[16] The best non-local estimation in the OOS turns out to be one that uses a window of length 500, much shorter than the total data available, making this also a type of local method. That method also significantly beats the benchmark model, with a GW statistic of -4.758, however this method is not included in the MCS.

the Fed Funds Rate (FFR) as a state variable: when combined with time it performs better than the benchmark, but when considered on its own the OOS average loss is not significantly lower than the benchmark, and is actually higher than using non-local QML on a shorter estimation window (500 or 250 observations). The other yield curve-based state variable (10Y-2Y) fares similarly when combined with time, and when considered on its own there is no bandwidth between zero and three standard deviations that is better than the non-local method in the validation sample, and so its optimal bandwidth is set to infinity; that is, when using 10Y-2Y as a state variable the optimal bandwidth is such that this conditioning variable is ignored.

[INSERT TABLE 1 ABOUT HERE]

To better understand the source of the improvement in performance of the best local method, Figure 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.

[INSERT FIGURE 2 ABOUT HERE]

In the supplemental appendix we consider a "local" analysis of the GARCH-X model, using

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 S1 shows that 12 methods significantly (at the 0.05 level) beat the improved benchmark GARCH-X QML method, which ranks 17^{th} out of the 26 competing methods. We find five 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 including an additional variable in the model, and 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:

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

By exploiting the information in high frequency data, this model has been widely found to outperform 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.

Table 2 presents results on the out-of-sample forecast performance of the various estimation methods. The benchmark method ranks 9^{th} out of the 13 estimators, and is significantly beaten, at the 0.05 level, by two local methods, based on VIX or time and VIX.[17] The latter of these is

[17] In the validation sample we find that when time is combined with RV, FFR or 10Y-2Y, the effective optimal bandwidths for the latter variables are infinite, and so these local methods reduce to one that just uses time as a state variable. In the OOS period, this means that there are apparently four methods tied for third place, though of course the latter three are redundant given the first, and so it is perhaps more correct to say that the benchmark model ranks 6^{th} out of the 10 *unique* competing methods.

the local method that is selected using the validation sample, and the GW statistic comparing this method to the benchmark is -2.66, strongly rejecting the benchmark in favor of the local estimator. 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]

[INSERT TABLE 2 ABOUT HERE]

To illustrate how local and non-local estimation leads to different forecasts, Figure 3 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 osberve that the local HAR produces forecasts that increase more quickly as market turbulence rose, and then decrease more quickly in the subsequent weeks.

[INSERT FIGURE 3 ABOUT HERE]

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 $\mathcal{F}_{t-1}$:

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

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

[18] Table S2 in the supplemental appendix presents results when the HAR-X model is taken as the baseline model. We find 21 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.

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 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):

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

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%.

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

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

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.[19] Instead, we estimate $[a, b]$ using the standardized residuals based on the standard QML GARCH series, and only localize the GARCH

[19] Table S.3 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.

model parameters. This leads to the GARCH-EDF model, Equation (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 the full estimation window, or windows of length 250, 500 or 1000 observations. 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, turns out to also perform best in the OOS period, and it significantly beats the benchmark, which is ranked 9^{th}, with a GW statistic of -3.23. Two other methods have significantly lower OOS than the benchmark, and both are local methods, using RV or VIX as state variables. The MCS contains four methods: the three local methods just discussed, as well as non-local estimation with a window of length 1000, though the latter of these does not significantly beat the benchmark method. Similar to the GARCH and HAR applications, we do not find that the yield curve-based state variables (FFR and 10Y-2Y) are helpful for forecasting these risk measures; in this application the optimal bandwidths for these state variables is found to be infinity.

[INSERT TABLE 3 ABOUT HERE]

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:

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

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 $\{\beta_{i,t}\}_{t=1}^{T}$, 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 $[\beta_{1,t}, \beta_{2,t}, \beta_{3,t}]$ 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 equations (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. We use the same state variable (and same bandwidth value) for all three AR(1) models, although that could be relaxed.[20] We additionally consider the usual, non-local, DNS model, estimated on the full sample, as well as

[20] 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.

windows of length 250, 500 and 1000 observations.

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.[21] 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, are humbling for the local methods: the two best methods are non-local OLS estimation using (relatively short) windows of 250 and 500 observations. 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. The best local method from the validation sample uses time and RV as state variables, and it significantly beats the benchmark method (GW statistic of -12.87), however it is significantly beaten by the non-local method with a window of 500 observations, with a t-statistic of 5.53.

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 8^{th} out of the 13 estimators, and it is significantly beaten by six alternative methods.[22] Comparing the best non-local method with the local method selected using the validation sample, which is one that uses time and VIX as state variables, we obtain a GW t-statistic of -6.91; strong evidence that the local method out-performs the non-local benchmark. The 95% model confidence set contains, effectively, just one estimator, the local method using time as the state variable.

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

[21] 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.

[22] In this application, the Federal Funds Rate turns out to be an uninformative state variable: the validation sample-optimal bandwidth is infinity, both when considered alone and when considered jointly with time. The other yield curve state vairable, 10Y-2Y, also adds nothing when combined with time. As the local method using time alone performs best in the OOS period, this leads to an apparent tie for first place, though naturally the second and third models add nothing beyond the first.

more difficult forecasting problems, alternative estimation methods like the local methods proposed here offer the possibility of yielding improved forecasts.

[INSERT TABLE 4 ABOUT HERE]

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. 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 and yield curve (h=1) applications, and VIX for the HAR, VaR-ES and yield curve (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 (2005). 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. For all five applications the state variable is either RV or VIX, and we see that the slope coefficient is positive in all of these cases, indicating that the local method does relatively worse when volatility is high. Only in the GARCH and VaR-ES applications, however, is the slope coefficient significantly different from zero.

[INSERT TABLE 5 ABOUT HERE]

To gain a more nuanced understanding of the relationship between OOS loss differences and the state variable, Figures 4 and 5 present a simple nonparametric kernel smooth of this relationship,

along with pointwise 95% confidence intervals.[23] 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 Figure 4 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 for RV above about 20% non-local QML shows some evidence of outperforming local QML.[24] Similar results hold for the VaR-ES comparison.

In Figure 5, as well as the middle panel of Figure 4, we see that the predicted OOS loss difference is almost constant in the state variable. For the one-day horizon yield curve forecasts, in the upper panel of Figure 5, we observe some evidence that the outperformance of the local method is particularly strong when volatility is low (the loss difference is more negative), but for RV above about 8% the relationship is approximately flat.

Finally, we use the recently-proposed "conditional superior predictive ability" (CSPA) test of Li *et al.* (2021) to test whether the non-local method has weakly lower expected loss across the entire support of the state variable:

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

as well the hypothesis where the inequality in equation (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 Figure 4. 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 dominating non-local estimation unconditionally, and outperforming

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

[24] 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 (2005) and Timmermann and Zhu (2021) for example. We do not pursue this extension here.

pointwise for low values of VIX as in Figure 4. 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.

[INSERT FIGURES 4 AND 5 ABOUT HERE]

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.

References

[1] Andersen, T.G., T. Bollerslev and F.X. Diebold, 2007, Roughing it up: Disentangling continuous and jump components in measuring, modeling and forecasting asset return volatility, *Review of Economics and Statistics*, 89(4), 701-720.

[2] Ang, A., G. Bekaert and M. Wei, 2007, Do macro variables, asset markets, or surveys forecast inflation better? *Journal of Monetary Economics*, 54(4), 1163-1212.

[3] Ang, A. and D. Kristensen, 2012, Testing conditional factor models, *Journal of Financial Economics*, 106, 132-156.

[4] Basel Committee on Banking Supervision, 2010, *Basel III: A Global Regulatory Framework for More Resiliant Banks and Banking Systems*, Bank for International Settlements. http://www.bis.org/publ/bcbs189.pdf.

[5] Beare, B.K., 2010, Copulas and temporal dependence, *Econometrica*, 78, 395-410.

[6] Blasques F., S.J. Koopman, M. Mallee, and Z. Zhang, 2016, Weighted maximum likelihood for dynamic factor analysis and forecasting with mixed frequency data, *Journal of Econometrics*, 193, 405-417.

[7] Bollerslev, T., 1986, Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics*, 31, 307–327.

[8] Brown, R.G., 1956, *Exponential smoothing for predicting demand.* Arthur D. Little Inc., Cambridge, Massachusetts.

[9] Capistran, C. and A. Timmermann, 2009, Forecast combination with entry and exit of experts, *Journal of Business & Economic Statistics*, 2009, 27, 429-440.

[10] Chen, X., and Y. Fan, 2006, Estimation of copula-based semiparametric time series models, *Journal of Econometrics*, 130, 307-335.

[11] Christoffersen, P., K. Jacobs, C. Ornthanalai and Y. Wang, 2008, Option valuation with long-run and short-run volatility components, *Journal of Financial Economics*, 90, 272-297.

[12] Cleveland, W.S. and S.J. Devlin, 1988, Locally weighted regression: An approach to regression analysis by local fitting, *Journal of the American Statistical Association*, 83, 596–610.

[13] Corsi, F., 2009, A simple approximate long-memory model of realized volatility, *Journal of Financial Econometrics*, 7(2), 174-196.

[14] Dendramis, Y., G. Kapetanios and M. Marcellino, 2020, A similarity-based approach for macroeconomic forecasting, *Journal of the Royal Statistical Society, Series A*, 183(3), 801-827.

[15] Diebold, F.X. and C. Li, 2006, Forecasting the term structure of government bond yields, *Journal of Econometrics*, 130, 337-364.

[16] Diebold, F.X. and R. Mariano, 1995, Comparing predictive accuracy, *Journal of Business and Economic Statistics*, 13, 253-265.

[17] Engle, R.F. and G.G.J. Lee, 1999, A permanent and transitory component model of stock return volatility, in *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W. J. Granger*. R.F. Engle and H. White, eds. Oxford University Press, pp. 475–97.

[18] Fan, J. Y. Wu and Y. Feng, 2009, Local quasi-likelihood with a parametric guide, *Annals of Statistics*, 37(6B), 4153-4183.

[19] Fan, J., M. Farmen and I. Gijbels, 1998, Local maximum likelihood estimation and inference, *Journal of the Royal Statistical Society, Series B*, 60(3), 591-608.

[20] Faust, J. and J.H. Wright, 2009, Comparing Greenbook and reduced form forecasts using a large realtime dataset, *Journal of Business & Economic Statistics*, 27(4), 468-479.

[21] Fissler, T. and J.F. Ziegel, 2016, Higher order elicitability and Osband's principle, *Annals of Statistics*, 44(4), 1680-1707.

[22] Giacomini, R. and H. White, 2006. Tests of conditional predictive ability, *Econometrica*, 74, 1545-1578.

[23] Giacomini, R. and B. Rossi, 2010, Forecast comparisons in unstable environments, *Journal of Applied Econometrics*, 25(4), 595-620.

[24] Giacomini, R. and G. Ragusa, 2014, Theory-coherent forecasting, *Journal of Econometrics*, 182, 145-155.

[25] Gneiting, T., 2011, Making and evaluating point forecasts, *Journal of the American Statistical Association*, 106, 746–762.

[26] Granger, C.W.J., 1969, Prediction with a generalized cost of error function, *OR*, 20(2), 199-207.

[27] Hansen, P.R. and A. Lunde, 2005, A forecast comparison of volatility models: Does anything beat a GARCH (1,1)? *Journal of Applied Econometrics*, 20(7), 873-889.

[28] Hansen, P.R., A. Lunde, and J.M. Nason, 2011, The model confidence set, *Econometrica*, 79(2), 453-497.

[29] Hansen, P.R. and E.-I. Dumitrescu, 2021, How should parameter estimation be tailored to the objective? *Journal of Econometrics*, forthcoming.

[30] Hu, F., 1997, The asymptotic properties of the maximum-relevance weighted likelihood estimators, *Canadian Journal of Statistics*, 25(1), 45-59.

[31] Hu, F. and J. V. Zidek, 2002, The weighted likelihood, *Canadian Journal of Statistics*, 30(3), 347-371.

[32] Inoue, A., L. Jin and B. Rossi, 2017, Rolling window selection for out-of-sample forecasting with time-varying parameters, *Journal of Econometrics*, 196, 55-67.

[33] Inoue, A., L. Jin, D. Pelletier, 2020, Local-linear estimation of time-varying-parameter GARCH models and associated risk measures, *Journal of Financial Econometrics*, 19(1), 202-234.

[34] Komunjer, I., 2013, Quantile prediction, in G. Elliott and A. Timmermann (eds), *Handbook of Economic Forecasting*, Volume 2, 961-994, Elsevier, Oxford.

[35] Kristensen, D. and A. Mele, 2011, Adding and subtracting Black-Scholes: A new approach to approximating derivative prices in continuous-time models, *Journal of Financial Economics*, 102, 390-415.

[36] Li, Q. and J.S. Racine, 2007, *Nonparametric Econometrics*, Princeton University Press, Princeton.

[37] Li, J., Z. Liao and R. Quaedvlieg, 2021, Conditional superior predictive ability, *Review of Economic Studies*, forthcoming.

[38] Manganelli, S., 2009, Forecasting with judgment, *Journal of Business & Economic Statistics*, 27(4), 553-563.

[39] Muth, J.F., 1960, Optimal properties of exponentially weighted forecasts, *Journal of the American Statistical Association*, 55(290) 299-306.

[40] Nelson, C.R. and A.F. Siegel, 1987, Parsimonious modeling of yield curve, *Journal of Business*, 60, 473-489.

[41] Nolde, N. and J.F. Ziegel, 2017, Elicitability and backtesting: Perspectives for banking regulation, *Annals of Applied Statistics*, 11(4), 1833-1874.

[42] Patton, A.J., 2020, Comparing possibly misspecified forecasts, *Journal of Business & Economic Statistics*, 38(4), 796-809.

[43] Patton, A.J., J.F. Ziegel and R. Chen, 2019, Dynamic semiparametric models for expected shortfall (and value-at-risk), *Journal of Econometrics*, 211(2), 388-413.

[44] Pesaran, M.H., A. Pick, and M. Pranovich, 2013. Optimal forecasts in the presence of structural breaks. *Journal of Econometrics*, 177, 134-152.

[45] Pettenuzzo, D., A. Timmermann and R. Valkanov, 2014, Forecasting stock returns under economic constraints, *Journal of Financial Economics*, 114, 517-553.

[46] Richter, S. and E. Smetanina, 2020, Forecast evaluation and selection in unstable environments, working paper, Chicago Booth.

[47] Tibshirani, R. and T. Hastie, 1987, Local likelihood estimation, J*ournal of the American Statistical Association*, 82(398), 559-567.

[48] Timmermann, A. and Y. Zhu, 2021, Monitoring forecasting performance, *Journal of Econometrics*, forthcoming.

[49] Weiss, A.A., 1996, Estimating time series models using the relevant cost function, *Journal of Applied Econometrics*, 11(5), 539-560.

[50] Zumbach, G., 2006, The RiskMetrics 2006 methodology, working paper, RiskMetrics Group, Geneva, Switzerland.

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

	Method details			*Forecast performance*		
Rank	**StateVar**	**Bwidth**	**Window**	**AvgLoss**	**GW stat**	**MCS**
1*	time,RV	0.9995,0.34	full	0.320	-10.316	✓
2	RV	0.37	full	0.325	-10.395	×
3	time,VIX	0.995,0.28	full	0.333	-6.195	×
4	VIX	0.32	full	0.349	-6.001	×
5	time	0.995	full	0.371	-5.427	×
6	-	-	500	0.375	-4.758	×
7	-	-	250	0.376	-2.817	×
8	time,10Y-2Y	0.9975,0.25	full	0.380	-3.449	×
9	time,FFR	0.9975,0.49	full	0.381	-3.855	×
10	-	-	1000	0.382	-4.494	×
11	FFR	1.81	full	0.400	-1.592	×
12	-	-	full	0.402	★	×
=12	10Y-2Y	∞	full	0.402	0.000	×

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. The fourth column reports the window of data used in estimation, where "full" implies the entire in-sample period (2737 observations). The penultimate column reports Giacomini-White t-statistics of each model relative to the benchmark method (marked with ★), which is taken as the non-local method using the full estimation window, 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: Out-of-sample forecast performance for HAR models

	Method details			*Forecast performance*		
Rank	**StateVar**	**Bwidth**	**Window**	**AvgLoss**	**GW stat**	**MCS**
1*	time,VIX	0.999,0.62	full	0.246	-2.655	✓
2	VIX	1.8	full	0.252	-4.610	×
3	time	0.995	full	0.252	-0.291	×
=3	time,RV	0.995,∞	full	0.252	-0.291	×
=3	time,FFR	0.995,∞	full	0.252	-0.291	×
=3	time,10Y-2Y	0.995,∞	full	0.252	-0.291	×
7	10Y-2Y	1.91	full	0.253	-1.318	×
8	RV	2.86	full	0.253	-0.362	×
9	-	-	full	0.253	★	×
10	-	-	500	0.253	0.046	×
11	FFR	2.3	full	0.253	0.922	×
12	-	-	250	0.255	0.642	×
13	-	-	1000	0.300	1.056	×

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. The fourth column reports the window of data used in estimation, where "full" implies the entire in-sample period (2737 observations). The penultimate column reports Giacomini-White t-statistics of each model relative to the benchmark method (marked with ★), which is taken as the non-local method using the full estimation window, 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 3: Out-of-sample forecast performance for VaR-ES models

	Method details			*Forecast performance*		
Rank	**StateVar**	**Bwidth**	**Window**	**AvgLoss**	**GW stat**	**MCS**
1*	time,VIX	0.9995,1.25	full	-3.869	-3.227	✓
2	RV	1.4	full	-3.868	-4.423	✓
3	VIX	1.24	full	-3.863	-2.013	✓
4	-	-	1000	-3.861	-0.627	✓
5	time	0.9975	full	-3.861	-0.593	×
=5	time,RV	0.9975,∞	full	-3.861	-0.593	×
=5	time,FFR	0.9975,∞	full	-3.861	-0.593	×
=5	time,10Y-2Y	0.9975,∞	full	-3.861	-0.593	×
9	-	-	full	-3.855	★	×
10	10Y-2Y	∞	full	-3.855	0.000	×
11	FFR	∞	full	-3.855	0.000	×
12	-	-	500	-3.844	0.581	×
13	-	-	250	-3.102	1.517	×

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 Equation (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. The fourth column reports the window of data used in estimation, where "full" implies the entire in-sample period (2737 observations). The penultimate column reports Giacomini-White t-statistics of each model relative to the benchmark method (marked with ★), which is taken as the non-local method using the full estimation window, 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 4: Out-of-sample forecast performance for yield curve models

	Method details			*Forecast performance*		
Rank	**StateVar**	**Bwidth**	**Window**	**AvgLoss**	**GW stat**	**MCS**
Panel A: One-day forecast horizon						
1	-	-	500	0.157	-9.499	✓
2	-	-	250	0.158	-5.071	✓
3	time,VIX	0.999,1.91	full	0.158	-11.618	×
4*	time, RV	0.9995,1.46	full	0.158	-12.868	×
5	time,10Y-2Y	0.999,1.21	full	0.158	-14.128	×
6	time,FFR	0.999,1.01	full	0.158	-10.490	×
7	time	0.999	full	0.158	-14.523	×
8	RV	1.43	full	0.158	-9.099	×
9	-	-	1000	0.158	-1.605	×
10	FFR	0.9	full	0.158	-2.721	×
11	VIX	1.84	full	0.158	-4.440	×
12	10Y-2Y	1.26	full	0.158	-5.881	×
13	-	-	full	0.158	★	×
Panel B: Twenty-day forecast horizon						
1	time	0.999	full	0.241	-6.542	✓
=1	time,FFR	0.999,∞	full	0.241	-6.542	✓
=1	time,10Y-2Y	0.999,∞	full	0.241	-6.542	✓
4	time,RV	0.999,2.18	full	0.242	-6.304	×
5*	time,VIX	0.9995,1.7	full	0.244	-6.911	×
6	VIX	1.4	full	0.248	-2.399	×
7	10Y-2Y	2.08	full	0.250	-0.422	×
8	-	-	full	0.250	★	×
=8	FFR	∞	full	0.250	0.000	×
10	RV	1.5	full	0.250	1.095	×
11	-	-	500	0.250	0.172	×
12	-	-	1000	0.253	1.364	×
13	-	-	250	0.262	2.567	×

Notes to Table 4: 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. The fourth column reports the window of data used in estimation, where "full" implies the entire in-sample period (2737 observations). The penultimate column reports Giacomini-White t-statistics of each model relative to the benchmark method (marked with ★), which is taken as the non-local method using the full estimation window, 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 5: Conditional comparisons of forecasting models

				Yield curve	
	GARCH	**HAR**	**VaR-ES**	**h=1**	**h=20**
Intercept	−0.082	−0.007	−0.014	−0.894	−29.593
(std. err.)	(0.008)	(0.003)	(0.004)	(0.069)	(4.282)
[t-stat]	[−10.316]	[−2.655]	[−3.227]	[−12.868]	[−6.911]
Slope	0.091	0.029	0.035	0.009	0.0716
(std. err.)	(0.009)	(0.025)	(0.017)	(0.144)	(0.719)
[t-stat]	[10.440]	[1.182]	[2.104]	[0.061]	[0.010]

Notes to Table 5: 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 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 and yield curve (h=1) application, VIX for the HAR and VaR-ES application, and 10Y-2Y for the yield curve (h=20) application.

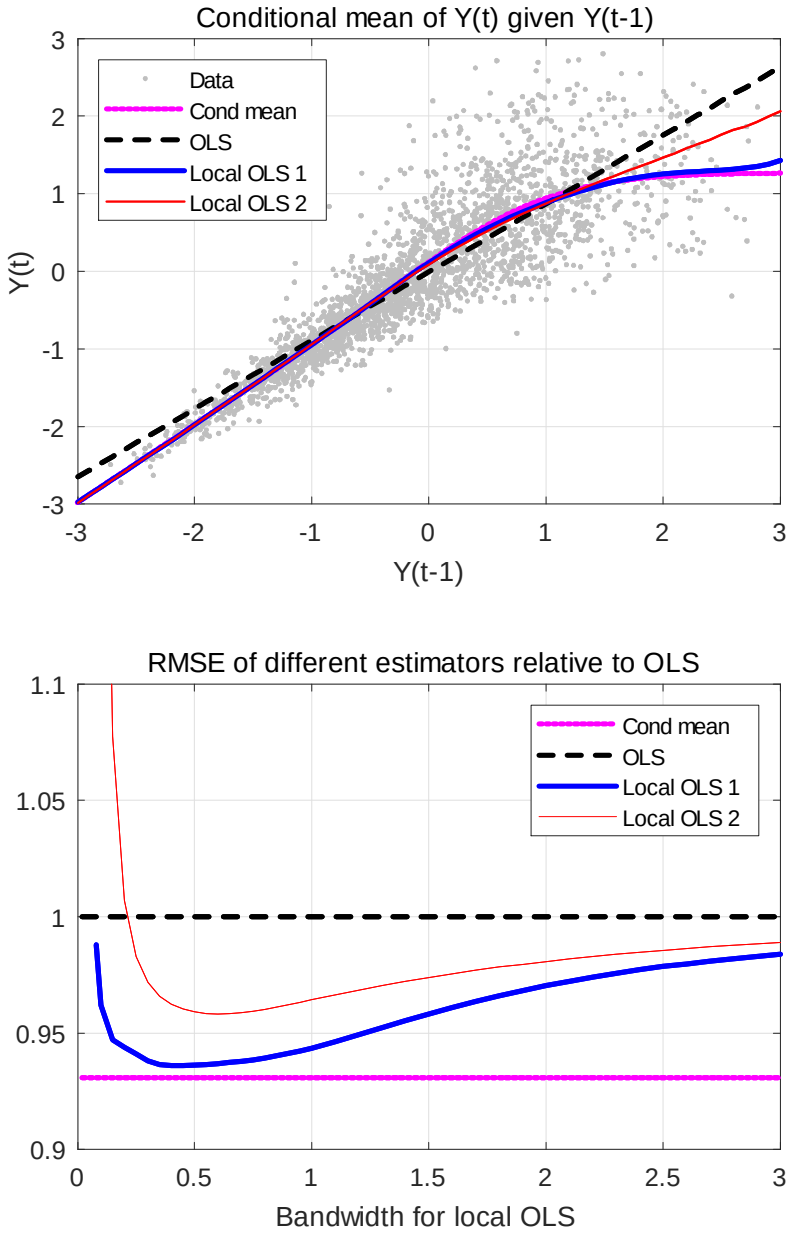

Figure 1: *The upper panel presents the expected value of Y_t given Y_{t-1} according to the DGP in equation (23), and estimates of this using a linear AR(1) estimated by OLS and local OLS with two different state variables: Y_{t-1} and Y_{t-2}. The lower panel presents the RMSE of the different estimators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

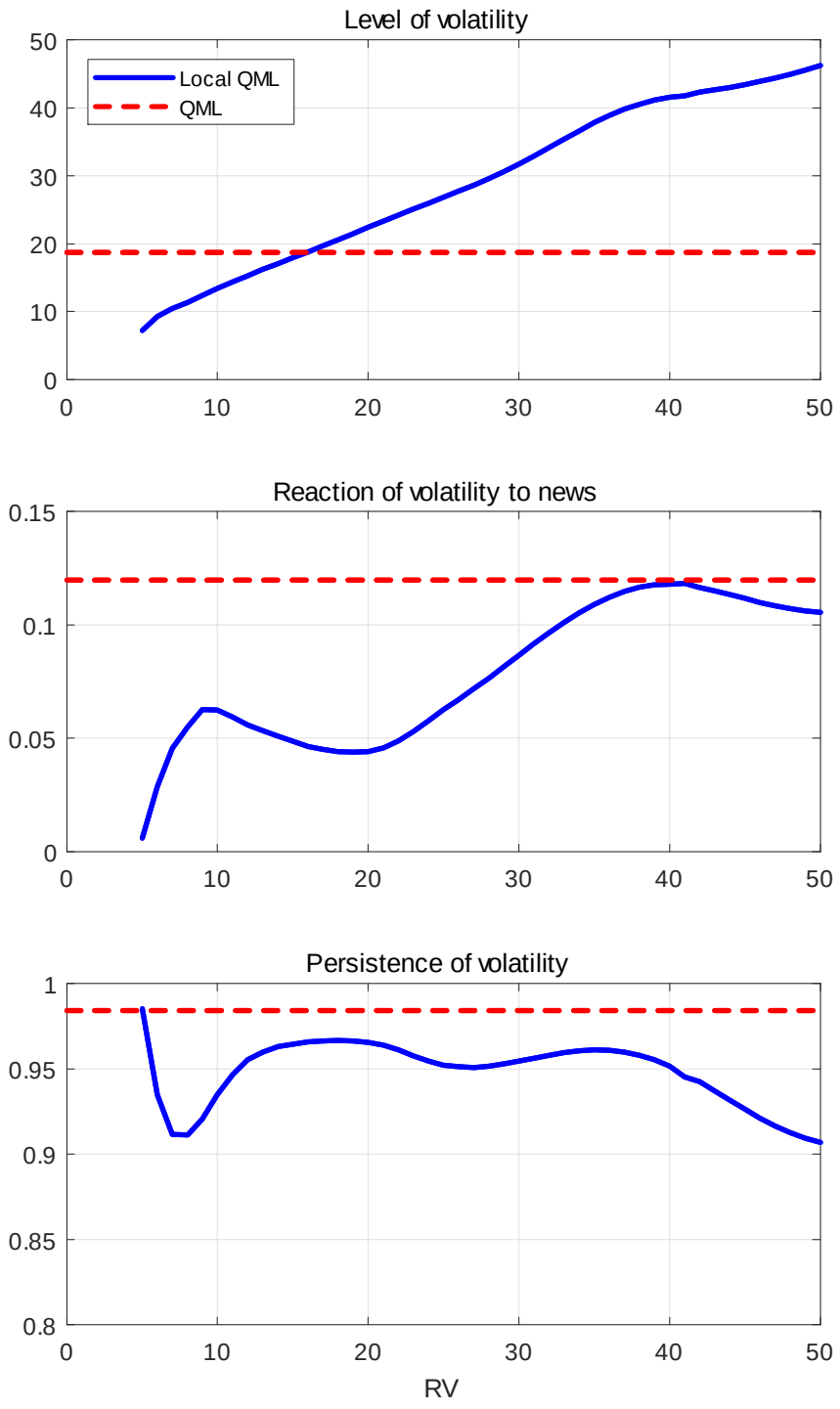

Figure 2: *This plot shows the local QML estimates of transformations of the GARCH(1,1) parameters* (ω, β, α) *as a function of realized volatility (RV). Also shown are the (non-local) QML parameter estimates. The upper, middle and lower panels plot* $\sqrt{\omega/(1-\alpha-\beta)}$, α, *and* $(\alpha+\beta)$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

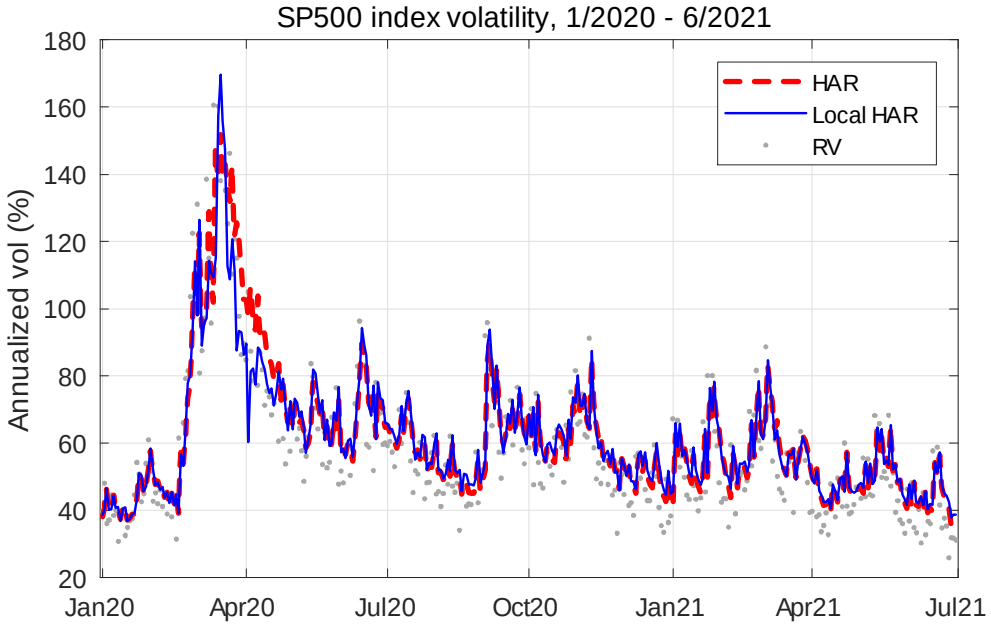

Figure 3: *This figure shows the predicted volatility from a HAR model estimated using local or non-local QML, along with realized volatility, over the last 18 months of the sample period.*

.

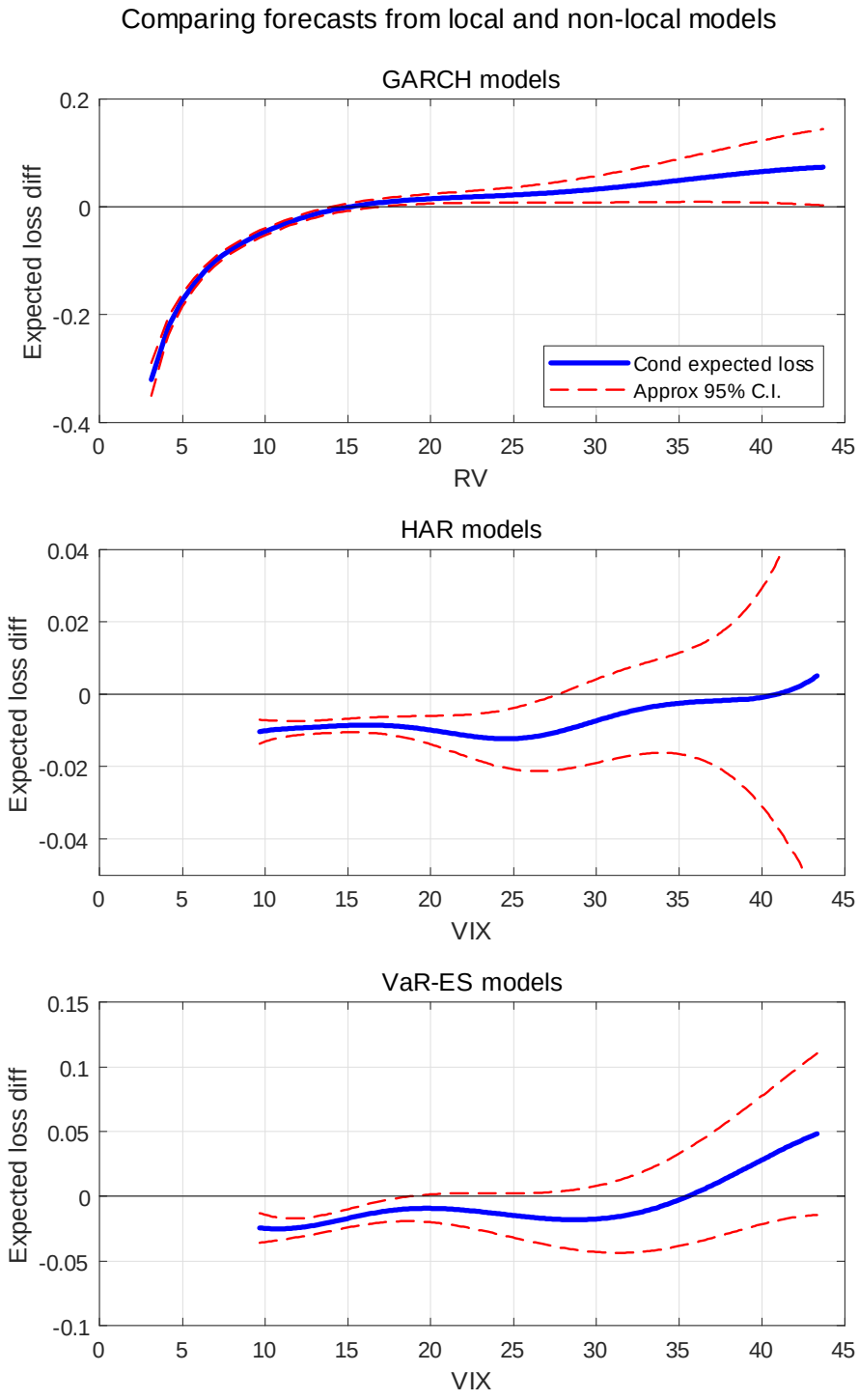

Figure 4: *This figure presents estimates of the expected out-of-sample loss differences from models estimated via local or non-local methods, conditional on realized volatility (top panel) or VIX (lower two panels). Positive loss differences indicate the non-local method is preferred.*

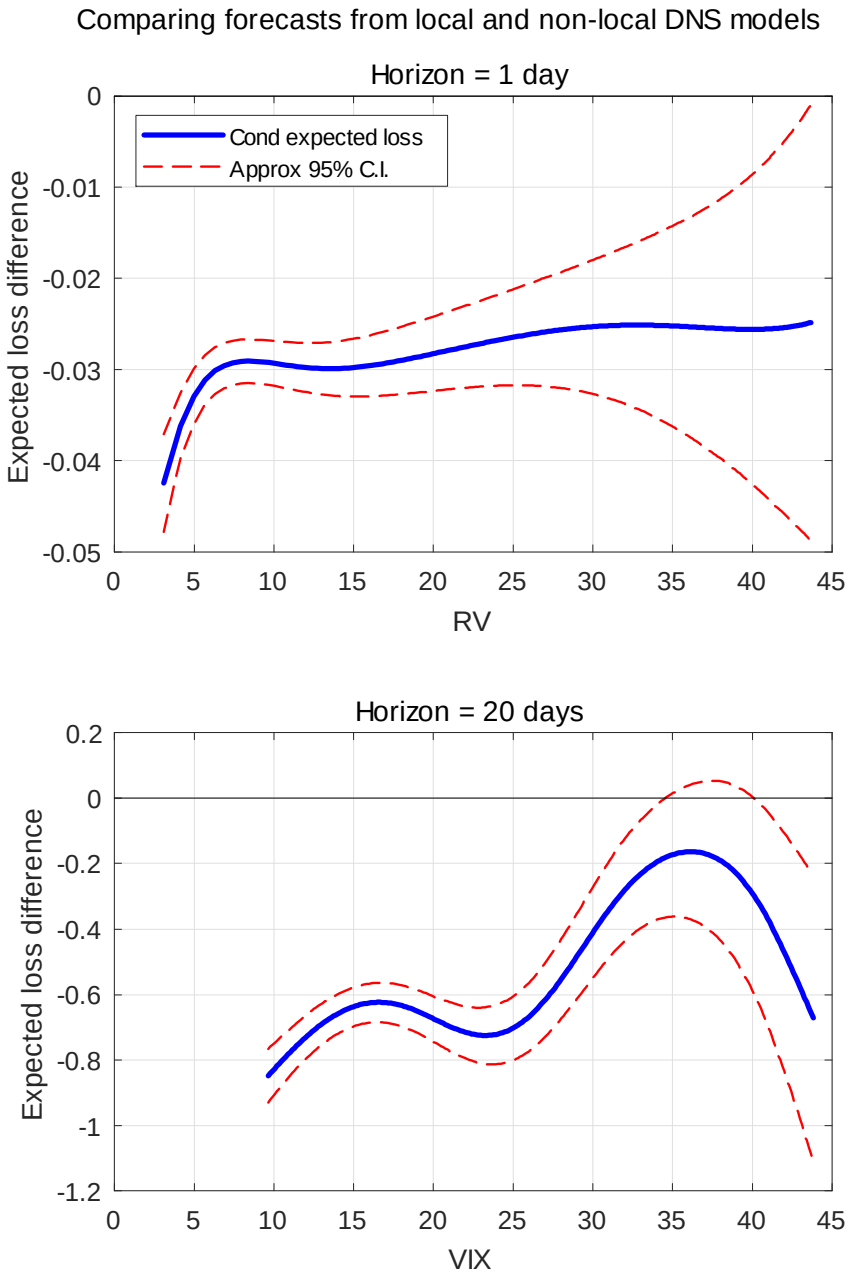

Figure 5: *This figure presents estimates of the expected out-of-sample loss difference of a dynamic Nelson-Siegel (DNS) model estimated via local OLS or non-local OLS, conditional on realized volatility (top panel) or VIX (bottom panel). Positive loss differences indicate the non-local method is preferred.*

Supplemental Appendix for

Better the Devil You Know: Improved Forecasts from Imperfect Models

by Dong Hwan Oh and Andrew J. Patton

12 October 2021

Table S1: Out-of-sample forecast performance for GARCH-X models

	Method details				*Forecast performance*		
Rank	**Model**	**StateVar**	**Bwidth**	**Window**	**AvgLoss**	**GW stat**	**MCS**
1	GARCH-X	time,RV	0.9999,0.4	full	0.293	-9.329	✓
2	GARCH-X	RV	0.41	full	0.294	-9.382	✓
3	GARCH-X	time,FFR	0.98,0.39	full	0.309	-5.017	✓
4	GARCH-X	time,VIX	0.98,0.89	full	0.313	-4.013	✓
5	GARCH-X	time	0.98	full	0.313	-4.653	✓
6	GARCH	time,RV	0.9995,0.34	full	0.320	-4.890	×
7	GARCH-X	-	-	250	0.324	-4.999	×
8	GARCH	RV	0.37	full	0.325	-4.239	×
9*	GARCH-X	time,10Y-2Y	0.9825,0.18	full	0.329	-1.828	×
10	GARCH	time,VIX	0.995,0.28	full	0.333	-2.294	×
11	GARCH-X	-	-	500	0.334	-5.330	×
12	GARCH-X	-	-	1000	0.335	-7.398	×
13	GARCH	VIX	0.32	full	0.349	-1.363	×
14	GARCH-X	VIX	2.63	full	0.351	-8.426	×
15	GARCH-X	10Y-2Y	0.26	full	0.358	-0.191	×
16	GARCH-X	FFR	0.44	full	0.359	-0.018	×
17	GARCH-X	-	-	full	0.359	★	×
18	GARCH	time	0.995	full	0.371	1.428	×
19	GARCH	-	-	500	0.375	2.049	×
20	GARCH	-	-	250	0.376	1.711	×
21	GARCH	time,10Y-2Y	0.9975,0.25	full	0.380	2.191	×
22	GARCH	time,FFR	0.9975,0.49	full	0.381	2.632	×
23	GARCH	-	-	1000	0.382	2.892	×
24	GARCH	FFR	1.81	full	0.400	4.623	×
25	GARCH	10Y-2Y	2.6	full	0.400	4.724	×
26	GARCH	-	-	full	0.402	4.844	×

Notes: This table presents measures of forecast performance over the out-of-sample period (January 2011 to June 2021) from GARCH and GARCH-X models estimated using either QML (non-local), or local QML. All GARCH-X models use VIX^2 as the extra variable. 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 third column and bandwidth parameter(s) from the fourth column, which are selected using the validation sample. The fifth column reports the window of data used in estimation, where "full" implies the entire in-sample period (2737 observations). The penultimate column reports Giacomini-White t-statistics of each model relative to the benchmark method (marked with ★), which is taken as the non-local method using the full estimation window, 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 S2: Out-of-sample forecast performance for HAR-X models

	Method details				*Forecast performance*		
Rank	**Model**	**StateVar**	**Bwidth**	**Window**	**AvgLoss**	**GW stat**	**MCS**
1	HAR-X	RV	0.79	full	0.232	-7.001	✓
2*	HAR-X	time,RV	0.9975,0.8	full	0.232	-6.843	✓
3	HAR-X	VIX	0.63	full	0.236	-6.722	×
4	HAR-X	time,10Y-2Y	0.9875,0.8	full	0.241	-6.085	×
5	HAR-X	time,VIX	0.9925,0.73	full	0.245	-5.723	×
6	HAR	time,VIX	0.999,0.62	full	0.246	-5.256	×
7	HAR-X	-	-	250	0.248	-5.576	×
8	HAR-X	time	0.995	full	0.248	-5.433	×
=8	HAR	time,RV	0.995,∞	full	0.251	-4.918	×
=8	HAR	time,10Y-2Y	0.995,∞	full	0.251	-4.918	×
=8	HAR	time,FFR	0.995,∞	full	0.252	-4.918	×
12	HAR	VIX	1.8	full	0.252	-4.829	×
13	HAR	time	0.995	full	0.252	-4.909	×
14	HAR	10Y-2Y	1.91	full	0.253	-4.789	×
15	HAR	RV	2.86	full	0.253	-4.743	×
16	HAR	-	-	full	0.253	-4.757	×
17	HAR	-	-	500	0.253	-4.785	×
18	HAR	FFR	2.3	full	0.253	-4.743	×
19	HAR	-	-	250	0.255	-4.742	×
20	HAR-X	time,FFR	0.99,0.32	full	0.263	-3.563	×
21	HAR-X	-	-	500	0.273	-3.097	×
22	HAR	-	-	1000	0.300	-0.533	×
23	HAR-X	-	-	1000	0.307	-0.782	×
24	HAR-X	-	-	full	0.325	★	×
25	HAR-X	10Y-2Y	1.96	full	0.351	2.564	×
26	HAR-X	FFR	1.62	full	0.372	3.734	×

Notes: This table presents measures of forecast performance over the out-of-sample period (January 2011 to June 2021) from HAR and HAR-X models estimated using either QML (non-local), or local QML. All HAR-X models use VIX^2 as the extra variable. 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 third column and bandwidth parameter(s) from the fourth column, which are selected using the validation sample. The fifth column reports the window of data used in estimation, where "full" implies the entire in-sample period (2737 observations). The penultimate column reports Giacomini-White t-statistics of each model relative to the benchmark method (marked with ★), which is taken as the non-local method using the full estimation window, 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 S.3: Out-of-sample forecast performance for GARCH-FZ models

	Method details			*Forecast performance*		
Rank	**StateVar**	**Bwidth**	**Window**	**AvgLoss**	**GW stat**	**MCS**
1	RV	1.96	full	-3.862	-4.136	✓
2	-	-	1000	-3.861	-0.619	✓
3	VIX	1.67	full	-3.860	-2.148	✓
4	10Y-2Y	2.72	full	-3.856	-1.249	×
5	-	-	full	-3.855	★	×
=5	FFR	∞	full	-3.855	0.000	×
7	time	0.995	full	-3.846	0.508	×
8	-	-	500	-3.836	0.876	×
9	time,RV	0.99,2.02	full	-3.830	1.071	×
10*	time,VIX	0.9925,1.21	full	-3.829	1.253	×
11	time,FFR	0.9925,2.24	full	-3.828	1.221	×
12	time,10Y-2Y	0.9925,1.04	full	-3.825	1.333	×
13	-	-	250	-3.812	1.308	×

Notes: This table presents measures of forecast performance over the out-of-sample period (January 2011 to June 2021) from GARCH-FZ models estimated using either M estimation or local M estimation and the FZ0 loss function in Equation (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. The fourth column reports the window of data used in estimation, where "full" implies the entire in-sample period (2737 observations). The penultimate column reports Giacomini-White t-statistics of each model relative to the benchmark method (marked with ★), which is taken as the non-local method using the full estimation window, 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.