Supplemental Appendix for

Better the Devil You Know: Improved Forecasts from Imperfect Models

by Dong Hwan Oh and Andrew J. Patton

November 2023

This supplemental appendix contains the following: Appendix SA.1 presents details for implementing the local M estimator and obtaining a forecast from the resulting estimated model. Appendix SA.2 presents results when the loss function used for estimation does not match the one used for forecast evaluation, and Appendix SA.3 presents a method for quantifying forecast improvements as a gain in effective sample size. Appendix SA.4 presents figures showing the sensitivity of the performance of forecasts from locally estimated models to the choice of bandwidth. This section also presents forecast comparison tables that include GARCH-X and HAR-X models, to see whether directly including the state variable in the model removes the gains from local estimation.

Section SA.1: Implementation details

This section presents details for implementing the local M estimator and obtaining a forecast from the resulting estimated model. We start with the case of local ordrinary least squares (OLS), which can be written as simple weighted least squares (and thus has an analytical solution), and then consider the general case of local M estimation.

The objective function for local OLS is:

$$\tilde{\theta}_{h,T}(s) = \arg\min_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - X_t'\theta\right)^2 K\left(s - S_{t-1}; h_T\right), \text{ for } s \in \text{Int}(\mathcal{S}) \tag{S.1}$$

For a given bandwidth, set $\omega_t(s) \equiv K(s - S_{t-1}; h_T)^{1/2}$, and define the re-weighted target variable $\tilde{Y}_t(s) \equiv \omega_t(s) Y_t$ and re-weighted regressors, $\tilde{X}_t(s) \equiv \omega_t(s) X_t$. Stack these into a $(T \times 1)$ vector $\ddot{\mathbf{Y}}_T(s)$ and a $(T \times p)$ matrix $\ddot{\mathbf{X}}_T(s)$. Then the solution to the above optimization problem is obtained as simply the OLS regression of $\tilde{Y}_t$ on $\tilde{X}_t$:

$$\tilde{\theta}_{h,T}(s) = \left(\ddot{\mathbf{X}}_T'(s)\ddot{\mathbf{X}}_T(s)\right)^{-1}\ddot{\mathbf{X}}_T'(s)\ddot{\mathbf{Y}}_T(s) \tag{S.2}$$

The above solution holds for any $s \in \text{Int}(\mathcal{S})$, but for out-of-sample forecasting we only need to evaluate the estimator at the prevailing value of the state variable. If we let the length of the estimation sample be denoted T, and the length of the prediction sample be denoted P, then the out-of-sample local OLS forecasts are obtained as:

$$\tilde{Y}_{t+1} \equiv X_t'\tilde{\theta}_{h,t}(S_t), \quad \text{for } t = T, T+1, ..., T+P-1 \tag{S.3}$$

where $\tilde{\theta}_{h,t}(S_t)$ is obtained using equation (S.2).

For local M estimation, including local QML, we do not generally have a closed-form expression

for the parameter estimate, $\tilde{\theta}_{h,T}(s)$. The out-of-sample local M forecasts are obtained as:

$$\hat{Y}_{t+1} \equiv g_t\left(\tilde{\theta}_{h^*,t}(S_t)\right), \quad \text{for } t = T, T+1, ..., T+P-1 \tag{S.4}$$

$$\text{where} \quad \tilde{\theta}_{h^*,t}(S_t) = \arg\min_{\theta\in\Theta} \frac{1}{t-1}\sum_{\tau=1}^{t-1} L\left(Y_{\tau+1}, g_\tau(\theta)\right) K\left(S_t - S_\tau; h^*\right) \tag{S.5}$$

and h^* is the optimal bandwidth from the validation sample (which runs for V periods finishing at time T). As starting values for the numerical optimization to obtain $\tilde{\theta}_{h^*,t}(S_t)$, we use the non-local parameter estimate, $\hat{\theta}_t$.

Section SA.2: Mis-matching loss functions for estimation and evaluation

Here we consider the case that the forecaster continues to use loss function L for forecast evaluation, but uses a loss function Q for estimation (either non-local or local). Such a situation can arise when a forecaster uses (local or non-local) quasi-maximum likelihood for estimation, while evaluation is conducted using a different loss function. For notational simplicity, we assume here that the parameter of the model, θ, is scalar.

Firstly, consider non-local estimation under Q, and define the estimator and its probability limit as:

$$\bar{\theta}_T \equiv \arg\min_{\theta\in\Theta} \frac{1}{T}\sum\nolimits_{t=1}^{T-1} Q\left(Y_{t+1}, g_t\left(\theta\right)\right) \tag{S.6}$$

$$\bar{\theta}^* \equiv \arg\min_{\theta\in\Theta} \mathbb{E}\left[Q\left(Y_{t+1}, g_t\left(\theta\right)\right)\right] \tag{S.7}$$

Next consider local estimation using the objective function Q, and define the estimator and its probability limit as:

$$\ddot{\theta}_{h,T}\left(s\right) = \arg\min_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} Q\left(Y_t, g_{t-1}\left(\theta\right)\right) K\left(s - S_{t-1}; h_T\right), \text{ for } s \in \text{Int}\left(\mathcal{S}\right) \tag{S.8}$$

$$\ddot{\theta}^*\left(s\right) \equiv \arg\min_{\theta\in\Theta} \mathbb{E}\left[Q\left(Y_{t+1}, g_t\left(\theta\right)\right)|S_t = s\right] \tag{S.9}$$

Comparing estimation using L and Q loss functions is related to the work of Hansen and Dumitrescu (2022), who study when full efficiency can be obtained by using a loss function for estimation that differs from that for evaluation. As in Hansen and Dumitrescu (2022), we distinguish between the case when L and Q are "coherent" versus "incoherent:" the former implies that $\bar{\theta}^* = \hat{\theta}^*$ and $\ddot{\theta}^*\left(s\right) = \tilde{\theta}^*\left(s\right)$, while the latter implies $\bar{\theta}^* \neq \hat{\theta}^*$ and $\ddot{\theta}^*\left(s\right) \neq \tilde{\theta}^*\left(s\right)$.

Comparison 1A: *Non-local estimation using Q vs. L, when L and Q are coherent*: In this case $\bar{\theta}_T$ (and $\hat{\theta}_T$) is $\sqrt{T}$-consistent for $\hat{\theta}^*$. We center the Taylor series expansion for $\bar{\theta}_T$ on $\hat{\theta}^*$ and find

$$\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\bar{\theta}_T\right)\right)\right] = \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\hat{\theta}^*\right)\right)\right] + \hat{H}^*\left(\bar{\theta}_T - \hat{\theta}^*\right)^2 + o_p\left(T^{-1}\right) \tag{S.10}$$

where $\hat{H}^*$ is defined in equation (16) of the main paper. The linear term in the Taylor series expansion vanishes since $\mathbb{E}[\partial L(Y_{T+1}, g_T(\hat{\theta}^*))/\partial\theta|\mathcal{F}_T] = 0$ from the definition of $\hat{\theta}^*$. The difference in expected loss is then

$$\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\bar{\theta}_T\right)\right)\right] - \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\hat{\theta}_T\right)\right)\right] = \hat{H}^*\left(\left(\bar{\theta}_T - \hat{\theta}^*\right)^2 - \left(\hat{\theta}_T - \hat{\theta}^*\right)^2\right) + o_p\left(T^{-1}\right) \tag{S.11}$$

and thus vanishes at rate $\mathcal{O}\left(T^{-1}\right)$. In finite samples, the sign of the out-of-sample loss difference is determined by which estimate is closer to the true parameter. In repeated samples, this corresponds to the estimator that is more efficient (i.e., has lower mean-squared estimation error), with that estimator generating lower out-of-sample loss.

Comparison 1B: *Non-local estimation using Q vs. L, when L and Q are incoherent*: This is the case that Hansen and Dumitrescu (2022) call "mindless estimation," and their Lemma 3 shows that the Q estimator (in our notation) is dominated by the L estimator asymptotically. We confirm and elaborate on this result below. First, consider a Taylor series expansion for $\bar{\theta}_T$ centered on $\bar{\theta}^*$:

$$\begin{aligned}
\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\bar{\theta}_T\right)\right)\right] &= \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\bar{\theta}^*\right)\right)\right] && \text{(S.12)}\\
&\quad + \frac{\partial \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\bar{\theta}^*\right)\right)\right]}{\partial\theta}\left(\bar{\theta}_T - \bar{\theta}^*\right) && \text{(S.13)}\\
&\quad + \bar{H}^*\left(\bar{\theta}_T - \bar{\theta}^*\right)^2 + o_p\left(T^{-1}\right) && \text{(S.14)}\\
\text{where} \quad \bar{H}^* &\equiv \frac{1}{2}\frac{\partial^2 \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\bar{\theta}^*\right)\right)\right]}{\partial\theta^2} && \text{(S.15)}
\end{aligned}$$

In this case the first-order term does not vanish as $\bar{\theta}^*$ is not the minimizer of $\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\cdot\right)\right)\right]$, rather it is the minimizer of $\mathbb{E}\left[Q\left(Y_{T+1}, g_T\left(\cdot\right)\right)\right]$. The sign of the first derivative is equal to $\text{sgn}\left(\bar{\theta}^* - \hat{\theta}^*\right)$. The difference in expected OOS loss is

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

The first term is positive, as $\hat{\theta}^*$ is the unique minimizer of $\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\cdot\right)\right)\right]$, and non-vanishing. The third term vanishes at rate $\mathcal{O}\left(T^{-1}\right)$, and reflects difference in estimation accuracy for L and Q estimation. The second term vanishes at rate $\mathcal{O}\left(T^{-1/2}\right)$ and is of indeterminate sign; it depends on the sign of the finite-sample bias in $\bar{\theta}_T$ and the sign of $\left(\bar{\theta}^* - \hat{\theta}^*\right)$. Consistent with Hansen and Dumitrescu (2022), the first term dominates, and reflects the penalty for the mismatch between estimation and evaluation loss functions.

Next we consider local estimation under L and Q.

Comparison 2A: *Local estimation using Q vs. L, when L and Q are coherent*: In this case $\ddot{\theta}_{h,T}\left(s\right)$ (and $\tilde{\theta}_{h,T}\left(s\right)$) is consistent for $\tilde{\theta}^*\left(s\right)$. For local estimation using Q the out-of-sample expected loss is:

$$\begin{aligned}\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\ddot{\theta}_{h,T}\left(S_T\right)\right)\right)\right] &= \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*\left(S_T\right)\right)\right)\right] \\ &\quad + \tilde{H}^*\left(S_T\right)\left(\ddot{\theta}_{h,T}\left(S_T\right) - \tilde{\theta}^*\left(S_T\right)\right)^2 + o_p\left(T^{-1+2\gamma}\right)\end{aligned} \tag{S.16}$$

where $\tilde{H}_T^*\left(S_T\right)$ is a Hessian term defined in equation (19) of the main paper. The difference in average out-of-sample loss is then

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

and vanishes at rate $\mathcal{O}\left(T^{-1+2\gamma}\right)$. In finite samples, the sign of the out-of-sample loss difference is determined by which estimate is more accurate (having smaller squared estimation error). In repeated samples this corresponds to the estimator that is more efficient, with that estimator generating lower out-of-sample loss.

Comparison 2B: *Local estimation using Q vs. L, when L and Q are incoherent*: In this case

$\ddot{\theta}^*(\cdot) \neq \tilde{\theta}^*(\cdot)$, and we center the Taylor series expansion for $\ddot{\theta}_{h,T}(S_T)$ around $\ddot{\theta}^*(S_T)$:

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

$$\text{where} \quad \ddot{H}^*(S_T) \equiv \frac{1}{2}\frac{\partial^2 \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\ddot{\theta}^*(S_T)\right)\right)\right]}{\partial \theta^2} \tag{S.18}$$

The first-order term does not vanish as $\ddot{\theta}^*(s)$ is not the minimizer of $\mathbb{E}\left[L\left(Y_{T+1}, g_T(\cdot)\right)|S_T=s\right]$, rather it is the minimizer of $\mathbb{E}\left[Q\left(Y_{T+1}, g_T(\cdot)\right)|S_T=s\right]$. Thus the difference in expected OOS loss is

$$\begin{aligned}&\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\ddot{\theta}_{h,T}(S_T)\right)\right)-L\left(Y_{T+1}, g_T\left(\tilde{\theta}_{h,T}(S_T)\right)\right)\right] \\ =\ &\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\ddot{\theta}^*(S_T)\right)\right)-L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*(S_T)\right)\right)\right] \\ &+\frac{\partial \mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\ddot{\theta}^*(S_T)\right)\right)\right]}{\partial \theta}\left(\ddot{\theta}_{h,T}(S_T)-\ddot{\theta}^*(S_T)\right) \\ &+\left(\ddot{H}^*(S_T)\left(\ddot{\theta}_{h,T}(S_T)-\ddot{\theta}^*(S_T)\right)^2-\tilde{H}^*(S_T)\left(\tilde{\theta}_{h,T}(S_T)-\tilde{\theta}^*(S_T)\right)^2\right)+o_p\left(T^{-1+2\gamma}\right)\end{aligned} \tag{S.19}$$

The first term on the right-hand side of equation (S.19) is positive, as $\tilde{\theta}^*$ is the minimizer of $\mathbb{E}\left[L\left(Y_{T+1}, g_T(\cdot)\right)\right]$, and it is non-vanishing. This term reflects a penalty for the mismatch between estimation and evaluation loss functions. The third term vanishes at rate $\mathcal{O}\left(T^{-1+2\gamma}\right)$ for some $\gamma \in (0, 1/2)$, and reflects the difference in estimation accuracy for local L and Q estimation. If Q is more accurate than L estimation, this term will be negative in repeated samples, thereby introducing a potential finite-sample trade-off between bias and variance. The second term vanishes at rate $\mathcal{O}\left(T^{-1/2+\gamma}\right)$ and is of indeterminate sign; it depends on the covariance between the estimation error in $\ddot{\theta}_{h,T}$ and the difference between $\ddot{\theta}^*$ and $\tilde{\theta}^*$, all evaluated at S_T.

In both the non-local and local comparisons, unless the loss functions L and Q are "coherent," in the terminology of Hansen and Dumitrescu (2022), forecasts obtained using a loss function for

estimation that differs from that for evaluation are expected to be worse than those obtained using the same loss function for both estimtion and evaluation. This is because the leading term in the difference in expected out-of-sample loss is strictly positive and non-vanishing. The next two terms are of indeterminate signs and vanish as the sample size grows. This is in contrast with the comparison of local and non-local forecasts in Section 2.3 of the main paper: there, the leading term is weakly negative (depending on the quality of the state variable and the accuracy of the forecasting model) and the next two terms are related to the estimation error in the local and non-local methods, and are known to be positive and negative respectively, vanishing at rates $T^{-1+2\gamma}$ for some $\gamma \in (0, 1/2)$ and T^{-1}. Essentially, it is the "quasi-coherency" of the local and non-local loss functions for estimation (local estimation under L nesting non-local estimation under L) that makes the comparison of out-of-sample forecast accuracy from local and non-local estimation intriguing.

Section SA.3: Quantifying the forecast improvements

While the improvements in out-of-sample forecast accuracy from using local models documented in the main paper are all statistically significant, interpreting the magnitude of these gains is more difficult. In this section we present one method for doing so, drawing on the trade-off between estimation error and goodness-of-fit that underlies comparing local and non-local models. We quantify the gains in forecast accuracy as a gain in sample size for estimating the benchmark model. Specifically, we determine the sample size for the benchmark model such that the improvement in fit achieved by going from the full-sample benchmark model to the full-sample local model, $\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\hat{\theta}_T\right)\right) - L\left(Y_{T+1}, g_T\left(\tilde{\theta}_T\left(S_T\right)\right)\right)\right] \equiv \hat{\mathcal{L}}_T - \tilde{\mathcal{L}}_T$, is the same as the improvement in going from a benchmark model estimated using a *hypothetical* sample of size aT, $a < 1$, to the full-sample benchmark model, $\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\hat{\theta}_{aT}\right)\right) - L\left(Y_{T+1}, g_T\left(\hat{\theta}_T\right)\right)\right] \equiv \hat{\mathcal{L}}_{aT} - \hat{\mathcal{L}}_T$.

To determine this equivalent sample size, consider a second-order Taylor series expansion of the expected loss for the benchmark model, as in Section 2.3 of the main paper, and assume that the parameter is a scalar for simplicity. By centering the expansion on the probability limit of the benchmark model parameter, the leading term drops out, and the linear term is zero by the first-order condition defining $\hat{\theta}^*$, leaving only the second-order terms:

$$\hat{\mathcal{L}}_{aT} - \hat{\mathcal{L}}_T = \hat{H}^*\left(\hat{\theta}_{aT} - \hat{\theta}^*\right)^2 - \hat{H}^*\left(\hat{\theta}_T - \hat{\theta}^*\right)^2 + o_p\left(T^{-1}\right) \tag{S.20}$$

where $\hat{H}^* \equiv \frac{1}{2}\partial^2 \mathbb{E}_T\left[L\left(Y_{T+1}, g_T\left(\hat{\theta}^*\right)\right)\right]/\partial\theta^2$ as in equation (16) of the main paper. Thus the difference in expected loss is approximately equal to the (weighted) difference in estimation error. Clearly the value of $\left(\hat{\theta}_T - \hat{\theta}^*\right)^2$ is unknown, but recalling that under standard regularity assumptions for M estimation we have $\sqrt{T}\left(\hat{\theta}_T - \hat{\theta}^*\right) \xrightarrow{D} N\left(0, \Sigma\right)$, we know the expected value of that term is Σ/T. Thus we find

$$\mathbb{E}\left[\hat{\mathcal{L}}_{aT} - \hat{\mathcal{L}}_T\right] \approx \frac{1-a}{a}\hat{H}^*\Sigma/T \tag{S.21}$$

We then set the above equation equal to the difference in expected loss from using the benchmark

and local models, and solve for a^* :

$$a^* = \frac{\hat{H}^*\Sigma/T}{\hat{H}^*\Sigma/T + \left(\hat{\mathcal{L}}_T - \tilde{\mathcal{L}}_T\right)} \tag{S.22}$$

If $\hat{\mathcal{L}}_T - \tilde{\mathcal{L}}_T > 0$ (the benchmark model has greater loss than local model) and $\hat{H}^*\Sigma/T > 0$ (there is non-zero estimation error in the benchmark model), then $a^* \in (0,1)$. If the local model is worse than the benchmark model, and so $\hat{\mathcal{L}}_T - \tilde{\mathcal{L}}_T < 0$, there are two cases to consider: If it is "not too much worse," in the sense that $0 < \tilde{\mathcal{L}}_T - \hat{\mathcal{L}}_T < \hat{H}^*\Sigma/T$, then we will find $a^* > 1$, with the value increasing in the outperformance of the non-local method. If the local method is "much worse," in the sense that $\tilde{\mathcal{L}}_T - \hat{\mathcal{L}}_T > \hat{H}^*\Sigma/T$, then the above expression for a^* does not make sense (it returns a negative value), and in such cases we set $a^* \to \infty$. We do not observe any such cases in our applications. The expression for a^* when θ is a vector is

$$a^* = \frac{\text{tr}\left(\hat{H}^*\Sigma\right)/T}{\text{tr}\left(\hat{H}^*\Sigma\right)/T + \left(\hat{\mathcal{L}}_T - \tilde{\mathcal{L}}_T\right)} \tag{S.23}$$

All of the terms in the expression for a^* in equation (S.23) are estimable: $\hat{\mathcal{L}}_T$ and $\tilde{\mathcal{L}}_T$ are simply the observed out-of-sample average losses for the benchmark and local models respectively. Σ is the usual asymptotic covariance matrix of a non-local M estimator, and can be estimated in the usual way:

$$\begin{aligned}
\hat{\Sigma} &= \hat{A}^{-1}\hat{B}\hat{A}^{-1} \\
\hat{A} &= \frac{1}{R}\sum\nolimits_{t=1}^{R} \frac{\partial^2 L\left(Y_{t+1}, g_t\left(\hat{\theta}_R\right)\right)}{\partial\theta\partial\theta'} \\
\hat{B} &= \frac{1}{R}\sum\nolimits_{t=1}^{R} \frac{\partial L\left(Y_{t+1}, g_t\left(\hat{\theta}_R\right)\right)}{\partial\theta} \frac{\partial L\left(Y_{t+1}, g_t\left(\hat{\theta}_R\right)\right)}{\partial\theta'}
\end{aligned} \tag{S.24}$$

$\hat{H}^*$ can be estimated simply as $\hat{A}/2$, leading to the simplification $\text{tr}\left(\hat{H}^*\Sigma\right) = \text{tr}\left(\hat{B}\hat{A}^{-1}\right)/2$.

The above approach can be used directly for all but one of our applications: the benchmark model for the VaR-ES application uses multi-step estimation, with a mix of QMLE, sample quantiles

and expected shortfall, and M estimation, complicating the estimation of Σ. For that application we instead estimate Σ using a block bootstrap (Politis and Romano, 1994) with an average block length of 250 days. That application also uses a loss function that is not differentiable, complicating the estimation of $\hat{H}^*$. That matrix corresponds to 1/2 of the D_0 matrix defined in equation (31) of Patton *et al.* (2019), and we use the estimator for that matrix given their Theorem 3.

Table S1 presents the out-of-sample loss for the benchmark non-local model and the best local model, across the five applications considered in this paper, along with Giacomini-White t-statistics. The first three rows of this table are taken from Tables 1 to 4 and are included for ease of comparison. The bottom row reports the "proportional equivalent sample size," a^*. We see that for the two volatility applications, local estimation is equivalent to having around 83 and 45 times more data respectively, indicating that the local versions of these models produce forecasts that are substantially better than the non-local benchmarks. In the VaR-ES application, the equivalent sample size is only about 1.5% smaller than the full sample, suggesting that while the gains from local estimation are statistically significant, they are relatively small in magnitude, consistent with the well-known difficulty of estimating tail parameters and evaluating tail forecasts. For one-step yield curve forecasting, local estimation leads to strongly statistically significant improvements, with a GW t-statistic of less than -9, but the effective increase in sample size is only about 16%, consistent with the benchmark forecasts already being very good. Finally, for 20-step-ahead yield curve forecasts local estimation is equivalent to having over four times the sample size, a meaningful improvement to match the strong statistical significance revealed by the GW t-statistic.

Table S1: Equivalent sample sizes for non-local models

				Yield curve	
	GARCH	**HAR**	**VaR-ES**	**h=1**	**h=20**
Benchmark OOS loss	0.414	0.254	-3.843	0.158	0.246
Best local OOS loss	0.315	0.238	-3.868	0.158	0.244
GW t-statistic	-12.302	-6.158	-1.987	-9.471	-6.966
Equiv sample (a^*)	0.012	0.022	0.986	0.838	0.234

Notes: This table presents the out-of-sample loss for the benchmark non-local model and the best local model, across the five applications considered in this paper, along with Giacomini-White t-statistics. The first three rows of this table are taken from Tables 1 to 4 and are presented here for ease of comparison. The bottom row reports the "proportional equivalent sample size," a^*, introduced above.

Section SA.4: Additional tables and figures

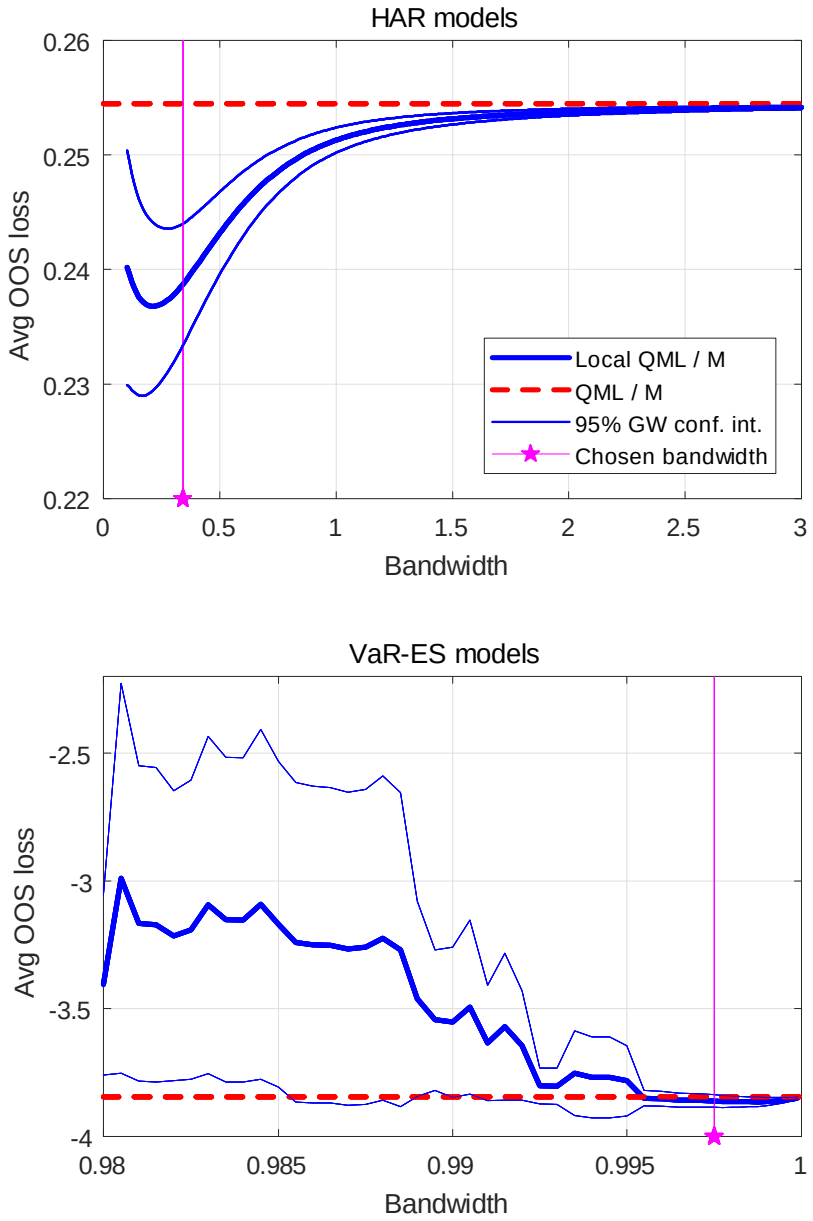

Figure S1: *This plot shows the average out-of-sample (OOS) loss using local QML estimates of a HAR model (upper panel) and local M-estimation of a VaR-ES model (lower panel), as a function of the bandwidth. The state variable for the HAR model is VIX. The state variable for the VaR-ES model is time. The dashed lines show the non-local average loss, and the thin solied lines show the 95% GW confidence interval on the difference between local and non-local average loss. The starred vertical lines show the bandwidths for each application chosen based on a validation sample separate from the OOS period.*

Table S2: Out-of-sample forecast performance for GARCH-X models

		Method details		*Forecast performance*		
Rank	**Model**	**StateVar**	**Bwidth**	**AvgLoss**	**GW stat**	**MCS**
1	GARCH	time,RV	0.9995,0.31	0.315	-6.473	✓
2	GARCH-X	RV	0.87	0.321	-13.820	✓
3	GARCH-X	time,RV	0.9999, 0.92	0.321	-13.721	✓
4	GARCH	RV	0.35	0.324	-5.236	✓
5*	GARCH-X	time,10Y-2Y	0.9775, 0.2	0.327	-2.834	✓
6	GARCH-X	time,VIX	0.96, 1.27	0.328	-2.805	✓
7	GARCH-X	time,FFR	0.9575, 0.5	0.331	-2.614	✓
8	GARCH-X	time	0.9575	0.333	-2.531	✓
9	GARCH	time,VIX	0.9999, 0.27	0.339	-6.521	×
10	GARCH	VIX	0.28	0.341	-6.211	×
11	GARCH-X	VIX	1.15	0.350	-8.914	×
12	GARCH-X	10Y-2Y	0.16	0.368	-1.813	×
13	GARCH	time	0.995	0.370	-0.466	×
14	GARCH-X	FFR	2.99	0.372	-10.435	×
15	GARCH-X	-	-	0.374	★	×
16	GARCH	time,FFR	0.9975, 0.33	0.377	0.250	×
17	GARCH	time,10Y-2Y	0.9975, 0.16	0.381	0.581	×
18	GARCH	10Y-2Y	0.11	0.396	2.019	×
19	GARCH	FFR	0.23	0.401	2.600	×
20	GARCH	-	-	0.414	3.977	×

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. 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.

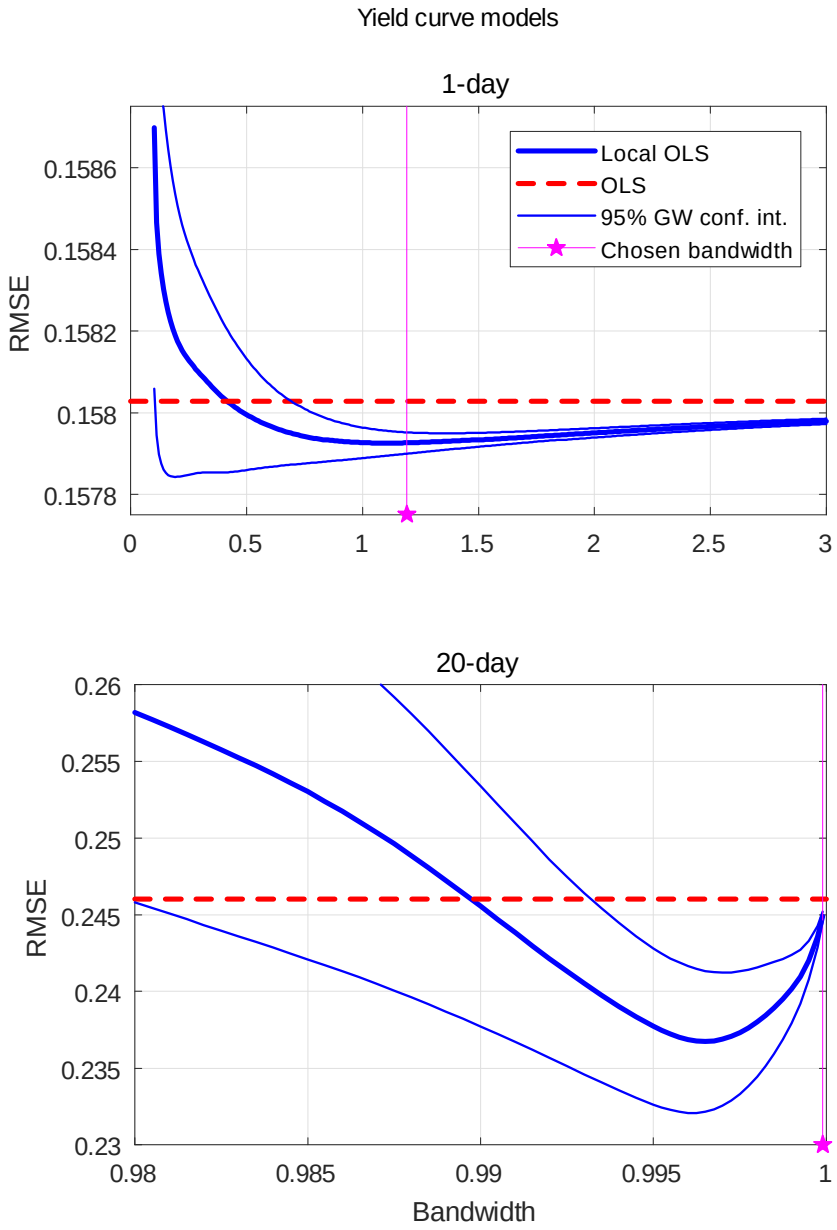

Figure S2: *This plot shows the out-of-sample RMSE using local OLS estimates of a yield curve models as a function of the bandwidth. The state variable for the 1-day horizon (upper plot) is realized variance. The state variable for the 20-day horizon (lower plot) is time. The dashed line is the (non-local) OLS RMSE, and the thin solied lines show the 95% GW confidence interval on the difference between local and non-local RMSE. The starred vertical lines show the bandwidths for each application chosen based on a validation sample separate from the OOS period.*

Table S3: Out-of-sample forecast performance for HAR-X models

		Method details		*Forecast performance*		
Rank	**Model**	**StateVar**	**Bwidth**	**AvgLoss**	**GW stat**	**MCS**
1	HAR-X	RV	0.47	0.229	-8.657	✓
2	HAR-X	time,RV	0.9975, 0.77	0.234	-7.620	✓
3	HAR-X	VIX	0.47	0.235	-8.177	×
4*	HAR-X	time,VIX	0.9999, 0.46	0.238	-7.913	×
5	HAR	time,VIX	0.9999, 0.35	0.238	-7.507	×
6	HAR	VIX	0.34	0.239	-7.484	×
7	HAR-X	time,FFR	0.99, 0.25	0.250	-5.609	×
8	HAR	time,10Y-2Y	0.9999, 2.5	0.254	-5.976	×
9	HAR	time,RV	0.9999, 3.05	0.254	-5.949	×
10	HAR	time	0.9999	0.254	-5.964	×
11	HAR	time,FFR	0.9999, 2.5	0.254	-5.929	×
12	HAR	10Y-2Y	2.68	0.254	-5.932	×
13	HAR	RV	3.11	0.254	-5.909	×
14	HAR	-	-	0.254	-5.921	×
15	HAR	FFR	2.51	0.255	-5.878	×
16	HAR-X	time	0.9999	0.331	-1.342	×
17	HAR-X	time,10Y-2Y	0.9999, 2.5	0.336	-0.387	×
18	HAR-X	-	-	0.340	★	×
19	HAR-X	10Y-2Y	2.12	0.352	1.452	×
20	HAR-X	FFR	2.01	0.373	3.356	×

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. 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 S4: Out-of-sample forecast performance for GARCH-FZ models

	Method details		*Forecast performance*		
Rank	**StateVar**	**Bwidth**	**AvgLoss**	**GW stat**	**MCS**
1	time	0.9995	-3.859	-2.731	✓
2	RV	2.51	-3.851	-5.587	✓
3	VIX	2.2	-3.850	-4.239	✓
4	FFR	2.03	-3.846	-1.873	×
5	-	-	-3.844	★	×
6	10Y-2Y	1.33	-3.843	0.343	×
7*	time,VIX	0.9925, 1.2	-3.838	0.325	×
8	time,RV	0.99, 1.6	-3.833	0.464	×
9	time,FFR	0.9925, 1.76	-3.830	0.686	×
10	time,10Y-2Y	0.9925, 2.28	-3.829	0.716	×

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. 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.

References

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

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

[3] Hansen, P.R. and E.-I. Dumitrescu, 2022, How should parameter estimation be tailored to the objective? *Journal of Econometrics*, 230, 535-558.

[4] Newey, W. K., and D. McFadden, 1994, Large Sample Estimation and Hypothesis Testing, in *Handbook of Econometrics* (Vol. 4), eds. R.F. Engle and D. McFadden, Amsterdam: North-Holland.

[5] 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.

[6] Politis, D.N. and J.P. Romano, 1994, The Stationary Bootstrap, *Journal of the American Statistical Association*, 89, 1303-1313.

[7] White, H., 1994, *Estimation, Inference and Specification Analysis*, Econometric Society Monographs No. 22, Cambridge University Press, Cambridge, U.K.

[8] White, H., 2001, *Asymptotic Theory for Econometricians* (2nd ed), San Diego: Academic Press.