Better the Devil You Know:

Improved Forecasts from Imperfect Models

Dong Hwan Oh
Federal Reserve Board

Andrew J. Patton
Duke University

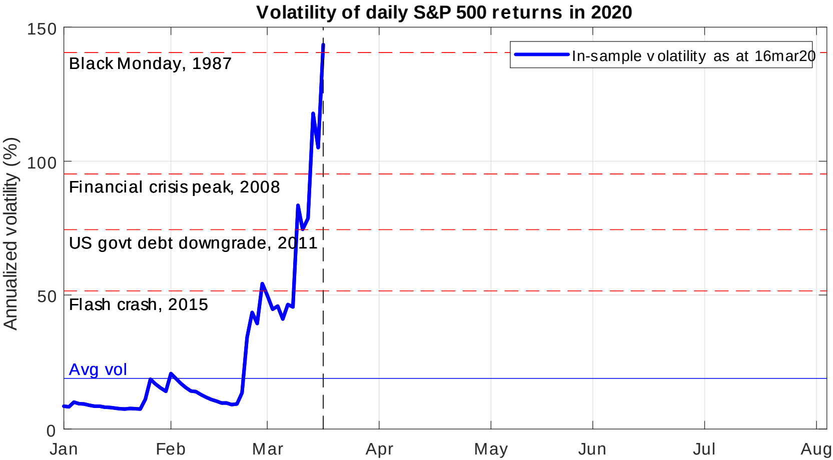
Dec 2023

Motivation

- Many important economic decisions are based on a forecasting model that is known to be good but *imperfect*.
- Why would an imperfect model be retained?
  - The model and its flaws are well-studied and understood
  - Institutional impediments to adopting a new model
  - Competitive environment too fast to change models
- Given a good but imperfect model, how can we improve the forecasts it produces?

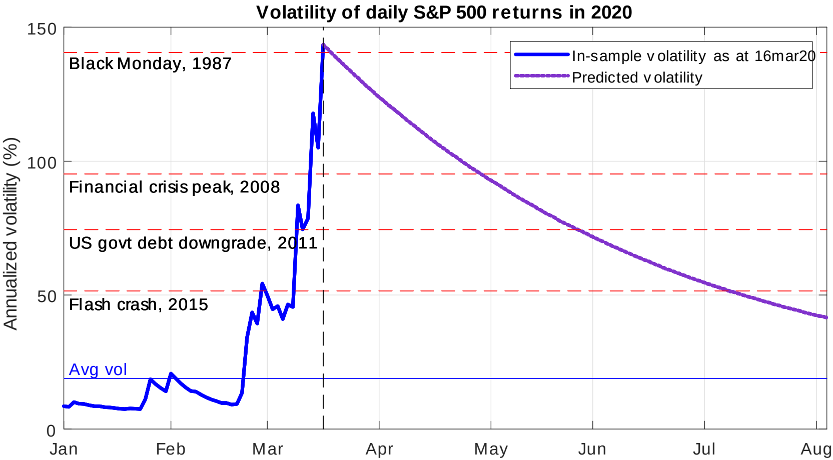
S&P 500 volatility in 2020

On 16 March 2020, volatility hit the highest it had been in over 30 years



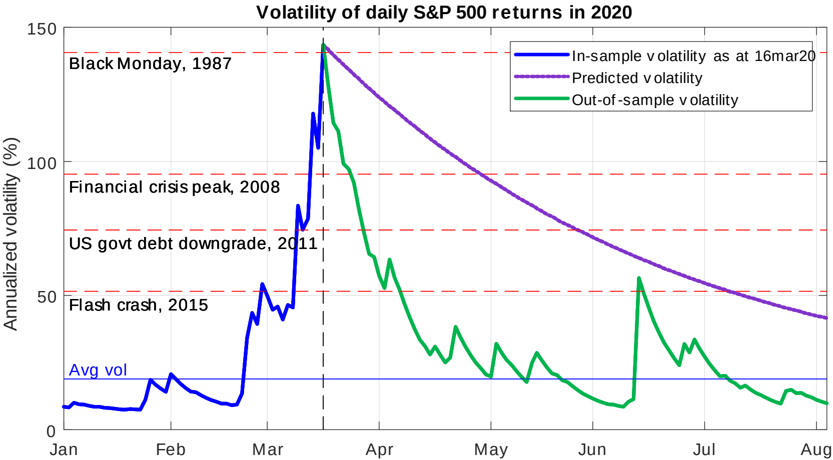
S&P 500 volatility in 2020

Long-range forecasts suggested that volatility would be high for a long time, but ...



S&P 500 volatility in 2020

Volatility did indeed mean revert faster than predicted by the GARCH model



Tilting the parameters of a misspecifed model

- We draw on info from a **state variable** that is informative about the misspecification of the model.
  - Eg, when the GARCH volatility is very high, we think mean reversion will be faster than the model implies
  - For models that are embedded in a decision-making process, such state variables are often available
- We propose estimating the forecasting model by emphasizing observations that are more similar to the forecast date.
  - Related: local OLS estimation and local MLE, see Tibshirani and Hastie (1987), Cleveland and Devlin (1988) and Fan et al. (1998).
  - Related: exponential smoothing, see Brown (1956) and Muth (1960).
- Importantly, we do not alter the model, only its parameters.

Contributions of this paper

- We consider local M estimation of a given parametric model for out-of-sample forecasting, drawing on past work.
  - Nests local OLS, local QML, etc.
- We theoretically analyze a bias-variance trade-off present in the local estimation framework, and obtain predictions for when such a method is likely to work well in practice.
  - Basline model cannot be too good; state variable cannot be too bad
- We apply the method to four distinct forecasting problems, and find significant improvements over the baseline methods in *almost* all cases.
  - We lose where we are predicted to lose...

Related literature

- **Local estimation**
  - Tibshirani and Hastie (1987, *JASA*), Cleveland and Devlin (1988, *JASA*), Fan, Farmen and Gijbels (1998, *JRSS*), Fan, Wu and Feng (2009, *AoS*), Dendramis, Kapetanios and Marcellino (2020, *JRSS*)
- **Smoothly-varying parameters**
  - Brown (1956, *book*), Muth (1960, *JASA*), Zumbach (2006, *wp*), Ang and Kristensen (2012, *JFE*), Inoue, Jin and Pelletier (2020, *JFEC*)
- **External information for forecasting**
  - Manganelli (2009, *JBES*), Giacomini and Ragusa (2014, *JoE*), Pettenuzzo, Timmermann and Valkanov (2014, *JFE*)

Outline

1. Motivation

2. Local estimation and a bias-variance trade-off

3. Applications
   - GARCH volatility forecasts using daily data
   - HAR volatility forecasts using high frequency data
   - VaR and ES forecasts
   - Yield curve forecasts

4. Conclusion

Outline

1. Motivation
2. **Local estimation and a bias-variance trade-off**
3. Applications
   - GARCH volatility forecasts using daily data
   - HAR volatility forecasts using high frequency data
   - VaR and ES forecasts
   - Yield curve forecasts
4. Conclusion

Target variable and target functional

- Our target variable is Y_{t+1} and our target functional is $g_t^\dagger$
  - Eg: $g_t^\dagger$ is the mean, a quantile, [VaR,ES], etc.
- L is a loss function that elicits the target functional:

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

  - Ruling out non-elicitable targets
- The baseline model is a parametric model for the target functional:

$$g\left(X_t, \theta\right) \equiv g_t\left(\theta\right)$$

which may or may not be correctly specified.

Estimation of the baseline model

- We assume the parameter of the baseline model is obtained via M estimation:

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

- Under standard conditions this has a well-defined probability limit:

$$\hat{\theta}^* \equiv \arg\min_{\theta \in \Theta} \mathbb{E}\left[L\left(Y_t, g_{t-1}\left(\theta\right)\right)\right]$$

- And converges at rate $\sqrt{T}$ to a Normal asymptotic distribution:

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

Incorporating information from a state variable

- Denote the forecaster's state variable as S_t
    - S_t must be $\mathcal{F}_t$-measurable, and may or may not be in the baseline model

Incorporating information from a state variable

- Denote the forecaster's state variable as S_t
  - S_t must be $\mathcal{F}_t$-measurable, and may or may not be in the baseline model
- We consider the estimator:

$$\tilde{\theta}_{h,T}(s) = \arg\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} \underbrace{L\left(Y_t, g_{t-1}(\theta)\right)}_{\text{loss function}} \times \underbrace{K\left(s - S_{t-1}; h_T\right)}_{\text{weighting function}}$$

where K is the kernel and $h_T \to 0$ is a bandwidth parameter.

Incorporating information from a state variable

- Denote the forecaster's state variable as S_t
  - S_t must be $\mathcal{F}_t$-measurable, and may or may not be in the baseline model
- We consider the estimator:

$$\tilde{\theta}_{h,T}(s) = \arg\min_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} \underbrace{L\left(Y_t, g_{t-1}(\theta)\right)}_{\text{loss function}} \times \underbrace{K\left(s - S_{t-1}; h_T\right)}_{\text{weighting function}}$$

where K is the kernel and $h_T \longrightarrow 0$ is a bandwidth parameter.

- Under regularity conditions, see e.g. Fan *et al.* (2009, *AoS*), the limit is:

$$\tilde{\theta}^{*}(s) \equiv \arg\min_{\theta\in\Theta} \mathbb{E}\left[L\left(Y_t, g_{t-1}(\theta)\right) | S_{t-1} = s\right]$$

Incorporating information from a state variable

- Denote the forecaster's state variable as S_t
  - S_t must be $\mathcal{F}_t$-measurable, and may or may not be in the baseline model
- We consider the estimator:

$$\tilde{\theta}_{h,T}(s) = \arg\min_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} \underbrace{L\left(Y_t, g_{t-1}(\theta)\right)}_{\text{loss function}} \times \underbrace{K\left(s - S_{t-1}; h_T\right)}_{\text{weighting function}}$$

where K is the kernel and $h_T \to 0$ is a bandwidth parameter.

- Under regularity conditions, see e.g. Fan *et al.* (2009, *AoS*), the limit is:

$$\tilde{\theta}^*(s) \equiv \arg\min_{\theta\in\Theta} \mathbb{E}\left[L\left(Y_t, g_{t-1}(\theta)\right) | S_{t-1} = s\right]$$

- With h_T shrinking at an appropriate rate, the estimator satisfies:

$$T^{1/2-\gamma}\left(\tilde{\theta}_{h,T}(s) - \tilde{\theta}^*(s)\right) = \mathcal{O}_p(1) \quad \text{for some } \gamma \in (0, 1/2)$$

The special case of correct specification

- If the baseline model is correctly specified, then

$$
\begin{aligned}
\exists\ \hat{\theta}^* \text{ s.t. } g_t^\dagger &\equiv \arg\min_{g \in \mathcal{G}} \mathbb{E}\left[L\left(Y_{t+1}, g\right)|\mathcal{F}_t\right] = g_t(\hat{\theta}^*) \\
\text{so } \mathbb{E}\left[L\left(Y_{t+1}, g_t(\hat{\theta}^*)\right)|\mathcal{F}_t\right] &\leq \mathbb{E}\left[L\left(Y_{t+1}, g\right)|\mathcal{F}_t\right] \ \ \forall\ g \in \mathcal{G}
\end{aligned}
$$

The special case of correct specification

- If the baseline model is correctly specified, then

$$\begin{aligned} \exists\ \hat{\theta}^* \text{ s.t. } g_t^{\dagger} &\equiv \arg\min_{g\in\mathcal{G}}\ \mathbb{E}\left[L\left(Y_{t+1},g\right)|\mathcal{F}_t\right]=g_t(\hat{\theta}^*) \\ \text{so } \mathbb{E}\left[L\left(Y_{t+1},g_t(\hat{\theta}^*)\right)|\mathcal{F}_t\right] &\leq \mathbb{E}\left[L\left(Y_{t+1},g\right)|\mathcal{F}_t\right]\ \ \forall\ g\in\mathcal{G} \end{aligned}$$

and since $S_t \in \mathcal{F}_t$, by the LIE we have

$$\mathbb{E}\left[L\left(Y_{t+1},g_t(\hat{\theta}^*)\right)|S_t\right]\leq\mathbb{E}\left[L\left(Y_{t+1},g\right)|S_t\right]\ \ \forall\ g\in\mathcal{G} \qquad (\bigstar)$$

The special case of correct specification

- If the baseline model is correctly specified, then

$$\exists\ \hat{\theta}^* \text{ s.t. } g_t^\dagger \equiv \arg\min_{g\in\mathcal{G}} \mathbb{E}\left[L\left(Y_{t+1},g\right)|\mathcal{F}_t\right] = g_t(\hat{\theta}^*)$$

$$\text{so } \mathbb{E}\left[L\left(Y_{t+1},g_t(\hat{\theta}^*)\right)|\mathcal{F}_t\right] \leq \mathbb{E}\left[L\left(Y_{t+1},g\right)|\mathcal{F}_t\right]\ \ \forall\ g\in\mathcal{G}$$

and since $S_t \in \mathcal{F}_t$, by the LIE we have

$$\mathbb{E}\left[L\left(Y_{t+1},g_t(\hat{\theta}^*)\right)|S_t\right] \leq \mathbb{E}\left[L\left(Y_{t+1},g\right)|S_t\right]\ \ \forall\ g\in\mathcal{G} \qquad (\bigstar)$$

- The local estimator satisfies:

$$\mathbb{E}\left[L\left(Y_{t+1},g_T(\tilde{\theta}^*\left(S_t\right))\right)|S_t\right] \leq \mathbb{E}\left[L\left(Y_{t+1},g_t\left(\theta\right)\right)|S_t\right]\ \ \forall\ \theta \qquad (\spadesuit)$$

The special case of correct specification

- If the baseline model is correctly specified, then

$$\exists\ \hat{\theta}^{*} \text{ s.t. } g_t^{\dagger} \equiv \arg\min_{g\in\mathcal{G}} \mathbb{E}\left[L\left(Y_{t+1},g\right)|\mathcal{F}_t\right]=g_t(\hat{\theta}^{*})$$

$$\text{so } \mathbb{E}\left[L\left(Y_{t+1},g_t(\hat{\theta}^{*})\right)|\mathcal{F}_t\right] \leq \mathbb{E}\left[L\left(Y_{t+1},g\right)|\mathcal{F}_t\right]\ \ \forall\, g\in\mathcal{G}$$

and since $S_t\in\mathcal{F}_t$, by the LIE we have

$$\mathbb{E}\left[L\left(Y_{t+1},g_t(\hat{\theta}^{*})\right)|S_t\right]\leq\mathbb{E}\left[L\left(Y_{t+1},g\right)|S_t\right]\ \ \forall\, g\in\mathcal{G} \qquad (\bigstar)$$

- The local estimator satisfies:

$$\mathbb{E}\left[L\left(Y_{t+1},g_T(\tilde{\theta}^{*}\left(S_t\right))\right)|S_t\right]\leq\mathbb{E}\left[L\left(Y_{t+1},g_t\left(\theta\right)\right)|S_t\right]\ \ \forall\,\theta \qquad (\spadesuit)$$

- Equations ($\bigstar$) and ($\spadesuit$) can only hold if $\tilde{\theta}^{*}\left(s\right)=\hat{\theta}^{*}$ **and is thus flat in** s.

A bias-variance trade-off

- Assume that L is differentiable and $\dim(\theta) = 1$.
- Then a second-order Taylor series expansion yields:

$$\mathbb{E}\left[L\left(Y_{T+1}, g_T(\tilde{\theta}_{h,T}(S_T))\right)\right] \approx \mathbb{E}\left[L\left(Y_{T+1}, g_T(\tilde{\theta}^*(S_T))\right)\right]$$

$$+ \frac{\partial \mathbb{E}\left[L\left(Y_{T+1}, g_T(\tilde{\theta}^*(S_T))\right)\right]}{\partial \theta}\left(\tilde{\theta}_{h,T}(S_T) - \tilde{\theta}^*(S_T)\right)$$

$$+ \frac{1}{2}\frac{\partial^2 \mathbb{E}\left[L\left(Y_{T+1}, g_T(\tilde{\theta}^*(S_T))\right)\right]}{\partial \theta^2}\left(\tilde{\theta}_{h,T}(S_T) - \tilde{\theta}^*(S_T)\right)^2$$

A bias-variance trade-off

- Assume that L is differentiable and $\dim(\theta) = 1$.
- Then a second-order Taylor series expansion yields:

$$\underbrace{\mathbb{E}\left[L\left(Y_{T+1}, g_T(\tilde{\theta}_{h,T}(S_T))\right)\right]}_{\text{loss using estimated params}} \approx \underbrace{\mathbb{E}\left[L\left(Y_{T+1}, g_T(\tilde{\theta}^*(S_T))\right)\right]}_{\text{loss using pop'n params}}$$

$$+ \underbrace{\frac{\partial \mathbb{E}\left[L\left(Y_{T+1}, g_T(\tilde{\theta}^*(S_T))\right)\right]}{\partial \theta}}_{=0 \text{ by FOC}} \left(\tilde{\theta}_{h,T}(S_T) - \tilde{\theta}^*(S_T)\right)$$

$$+ \underbrace{\frac{1}{2} \frac{\partial^2 \mathbb{E}\left[L\left(Y_{T+1}, g_T(\tilde{\theta}^*(S_T))\right)\right]}{\partial \theta^2}}_{\equiv \tilde{H}_T^* > 0} \underbrace{\left(\tilde{\theta}_{h,T}(S_T) - \tilde{\theta}^*(S_T)\right)^2}_{= \mathcal{O}_p(T^{-1+2\gamma}) \geq 0}$$

A bias-variance trade-off

- And similarly for the usual estimator:

$$\underbrace{\mathbb{E}\left[L\left(Y_{T+1}, g_T(\hat{\theta}_T)\right)\right]}_{\text{loss using estimated params}} \approx \underbrace{\mathbb{E}\left[L\left(Y_{T+1}, g_T(\hat{\theta}^*)\right)\right]}_{\text{loss using pop'n params}}$$

$$+ \underbrace{\frac{\partial \mathbb{E}\left[L\left(Y_{T+1}, g_T(\hat{\theta}^*)\right)\right]}{\partial \theta}}_{=0 \text{ by FOC}} \left(g_T(\hat{\theta}_T) - g_T(\hat{\theta}^*)\right)$$

$$+ \underbrace{\frac{1}{2}\frac{\partial^2 \mathbb{E}\left[L\left(Y_{T+1}, g_T(\hat{\theta}^*)\right)\right]}{\partial \theta^2}}_{\equiv \hat{H}_T^* > 0} \underbrace{\left(g_T(\hat{\theta}_T) - g_T(\hat{\theta}^*)\right)}_{=\mathcal{O}_p(T^{-1}) \geq 0}$$

- Note the order of the last term.

A bias-variance trade-off

- Finally, take expectations and consider the difference between the OOS losses:

$$\underbrace{\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}_{h,T}\left(S_T\right)\right)\right) - L\left(Y_{T+1}, g_T\left(\hat{\theta}_T\right)\right)\right]}_{\text{Actual diff in OOS loss}} \qquad \text{sign?}$$

$$\approx \underbrace{\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*\left(S_T\right)\right)\right) - L\left(Y_{T+1}, g_T\left(\hat{\theta}^*\right)\right)\right]}_{\text{Diff in OOS loss using pop'n params}} \qquad \leq 0 \text{ (bias } \downarrow)$$

$$+ \qquad \underbrace{\mathcal{O}_p\left(T^{-1+2\gamma}\right)}_{\text{Estimation error of local estimator}} \qquad > 0 \text{ (variance } \uparrow)$$

A bias-variance trade-off

- Finally, take expectations and consider the difference between the OOS losses:

$$\underbrace{\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}_{h,T}\left(S_T\right)\right)\right) - L\left(Y_{T+1}, g_T\left(\hat{\theta}_T\right)\right)\right]}_{\text{Actual diff in OOS loss}} \qquad \text{sign?}$$

$$\approx \underbrace{\mathbb{E}\left[L\left(Y_{T+1}, g_T\left(\tilde{\theta}^*\left(S_T\right)\right)\right) - L\left(Y_{T+1}, g_T\left(\hat{\theta}^*\right)\right)\right]}_{\text{Diff in OOS loss using pop'n params}} \qquad \leq 0 \text{ (bias } \downarrow\text{)}$$

$$+ \quad \underbrace{\mathcal{O}_p\left(T^{-1+2\gamma}\right)}_{\text{Estimation error of local estimator}} \qquad > 0 \text{ (variance } \uparrow\text{)}$$

Q: When is the improved fit unlikely to outweigh the increased estimation error?

Empirical predictions from the bias-variance analysis I

1. **Correctly specified models**. In this case we know

$$\tilde{\theta}^*(s) = \hat{\theta}^* \ \forall \ s$$

and so we have

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

- No improvement in fit from using local estimation
- Increased local estimation error causes 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 improvements are more than offset by estimation error is increased.

Empirical predictions from the bias-variance analysis II

2. **Bad state variables**. If the scores of the usual estimator are *mean independent* of the state variable S_t:

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

   - Then local estimation's FOC is satisfied when $\tilde{\theta}^*(S_t) = \hat{\theta}^*$
   - And so a bad state variable leads to $\tilde{\theta}^*(s)$ being flat in s. (Same outcome as in the correctly-specified case, but from a different source.)
   - No improvement in fit, and a loss from increased estimation error.

★ More generally, when the state variable is only weakly informative the gains from local estimation are lower, and the possibility that any gains are more than offset by estimation error is increased.

A stylized example

- Consider a nonlinear AR(1) process with standard Normal marginal distributions and a Clayton copula linking adjacent realizations:

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

where Φ is a standard Normal CDF, κ is the Clayton copula parameter.

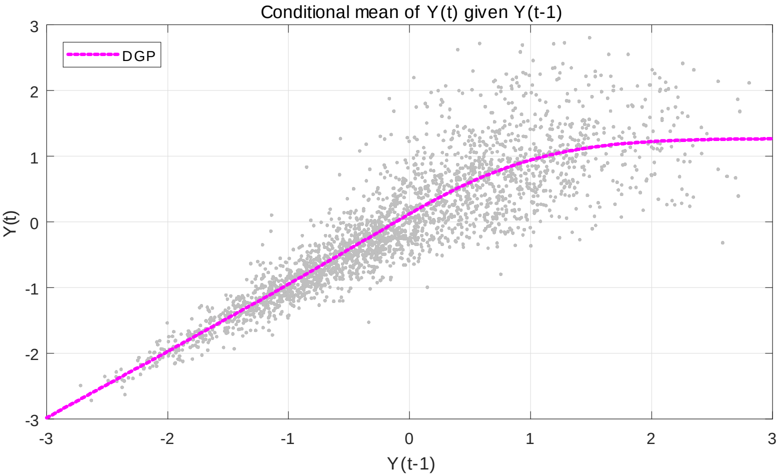
  - E.g., see Chen and Fan (2006, *JoE*) and Beare (2010, *ECMA*)
  - We set $\kappa = 5$ which implies first-order autocorrelation of about 0.85, and consider an estimation sample of $T = 1000$.

- Model is a linear AR(1)

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + e_t$$

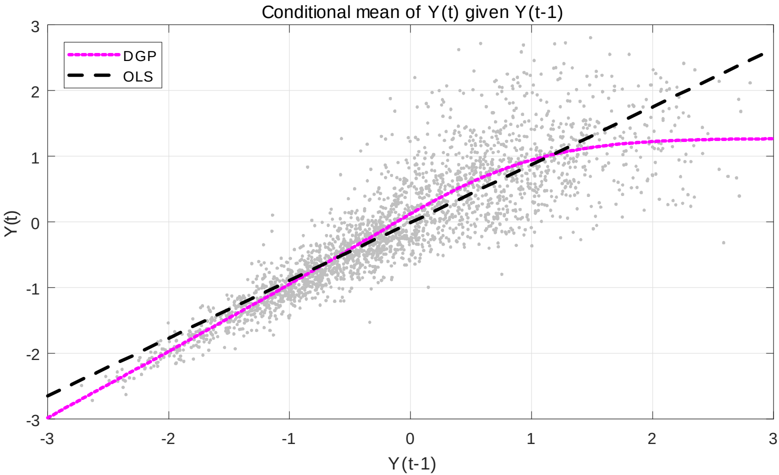
A stylized example

True conditional mean function is concave



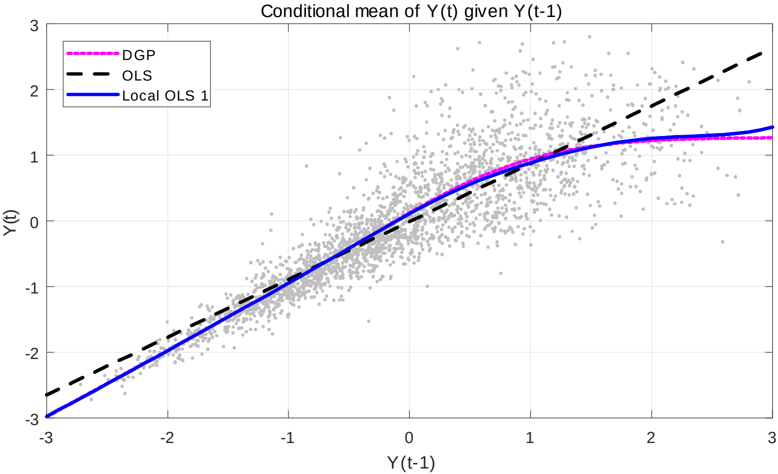
A stylized example

OLS provides best linear approximation



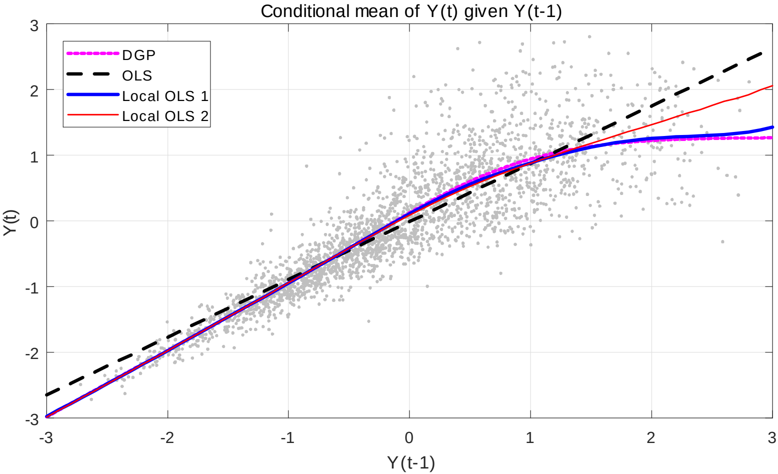
A stylized example

Local estimation using lagged Y as state variable is almost perfect



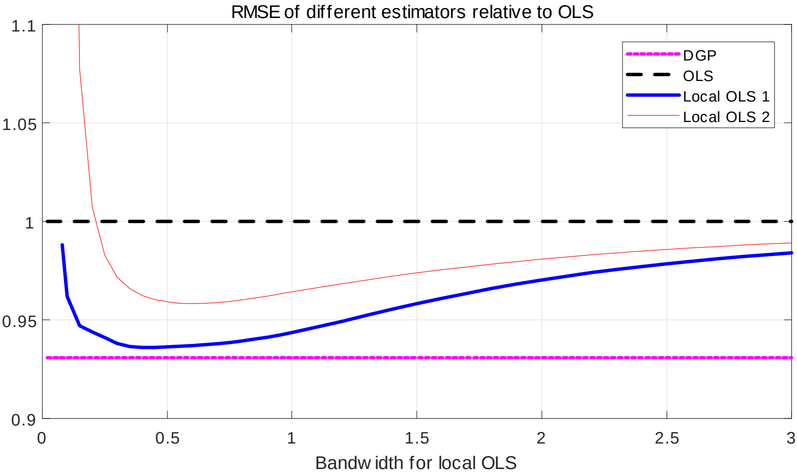
A stylized example

Local estimation using second lag of Y is (quite) good but not perfect



A stylized example

Bandwidth selection is important, especially with an imperfect state variable



Outline

1. Motivation

2. Local estimation and a bias-variance trade-off

3. **Applications**

   - GARCH volatility forecasts using daily data
   - HAR volatility forecasts using high frequency data
   - VaR and ES forecasts
   - Yield curve forecasts

4. Conclusion

Sample period and state variables

- Our sample period is Jan 2000 – June 2021, so $T \approx 5000$
  - Estimation sample is 2000-2010
    - "Training" sample is 2000-2005
    - "Validation" sample is 2006-2010
  - Out-of-sample period is 2011-2021
- We consider time and four stochastic state variables
  1. 5-minute realized volatility of S&P 500 index
  2. VIX (option-based volatility index)
  3. Fed Funds Rate
  4. 10-year minus 2-year Treasury yield
- We also consider 4 bivariate state variables using time & each of the above
  - Total of 9 state variables

Kernels and bandwidths

- For stochastic state variables we use a Gaussian kernel:

$$K_G\left(x;h\right)=\exp\left\{-\frac{x^2}{2h^2}\right\}$$

  - We consider values for h ranging from $0.01\sigma_S$ to $3\sigma_S$, where $\sigma_S^2=V\left[S_t\right]$.

- For time we use an exponential kernel:

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

  - We consider values for λ ranging from 0.98 to 0.9999.

- As $h\rightarrow\infty$ or $\lambda\rightarrow 1$ local estimation reduces to non-local estimation
- We estimate the models on the first half of the estimation sample and find the optimal bandwidth using the second half. **No look-ahead bias**.

Non-local estimation and forecast comparisons

- For non-local estimation we consider estimation windows of length 250, 500, 1000 and 2500 observations.

Non-local estimation and forecast comparisons

- For non-local estimation we consider estimation windows of length 250, 500, 1000 and 2500 observations.
  - Short estimation windows can be interpreted as a form of local estimation, where the state variable is time and the kernel is one-sided rectangular.
  - Despite this, we label these methods as “non local,” and treat them as part of the set of benchmark methods.

Non-local estimation and forecast comparisons

- For non-local estimation we consider estimation windows of length 250, 500, 1000 and 2500 observations.
  - Short estimation windows can be interpreted as a form of local estimation, where the state variable is time and the kernel is one-sided rectangular.
  - Despite this, we label these methods as "non local," and treat them as part of the set of benchmark methods.
- Forecast comparisons:

Non-local estimation and forecast comparisons

- For non-local estimation we consider estimation windows of length 250, 500, 1000 and 2500 observations.
  - Short estimation windows can be interpreted as a form of local estimation, where the state variable is time and the kernel is one-sided rectangular.
  - Despite this, we label these methods as "non local," and treat them as part of the set of benchmark methods.
- Forecast comparisons:
  - Giacomini-White (2006, *ECMA*) tests for pairwise comparisons with the OOS best non-local method.
  - Model confidence sets (Hansen *et al.*, 2011, *ECMA*) to compare all methods jointly.
  - Conditional comparisons, using GW (2006) and Li *et al.* (2021, *REStud*)

GARCH forecasts

- The GARCH model of Bollerslev (1986) is a very popular model for forecasting asset return volatility
  - 32,464 Google scholar citations as of this morning
- Assuming the conditional mean is zero, the model is:

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

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

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

- The local method minimizes weighted QLIKE loss.

GARCH forecasts

The benchmark non-local method is ranked (equal) last out of 13

	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	×

GARCH forecasts

The benchmark non-local method is signif beaten by all but two local methods

	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	×

GARCH forecasts

Yield curve variables are poor state variables in this application

	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	×

GARCH forecasts

The best non-local method is ranked 6th out of 13 (and is quite "local")

	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	×

GARCH forecasts

The best non-local method is not in the MCS

	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	×

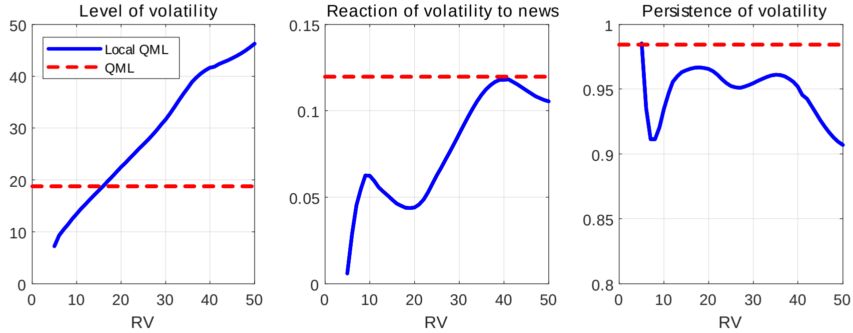
GARCH forecasts

The best local method in the validation sample also performs best out-of-sample

	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	×

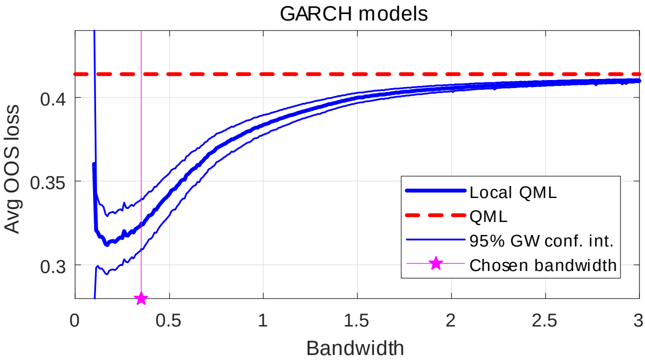
Local GARCH parameters

RV info enters through the level. Persistence drops for RV>35.



Sensitivity to choice of bandwidth

Local GARCH forecasts beat non-local forecasts for a range of bandwidths



HAR forecasts

- The HAR model of Corsi (2009) is a popular model for high frequency asset return volatility:

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

- We consider estimating this model either via standard OLS (the benchmark) or local OLS, conditioning on the same state variables as in the GARCH application.

HAR forecasts

Best non-local estimator ranked 9th; Best V-sample local estimator beats with t-stat of -2.7

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

HAR forecasts

Four models "tied" for 3rd place, but really the second state variable is just redundant.

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

HAR and Local HAR volatility forecasts in 2020-21

Local HAR forecasts recover more quickly after the shock of March 2020

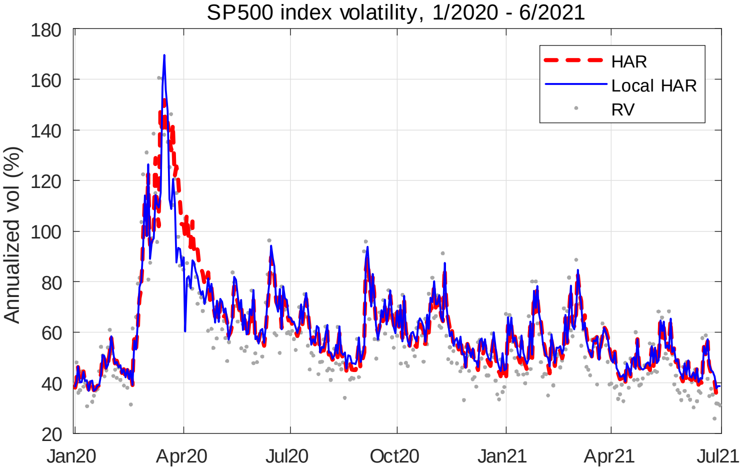


Figure:

VaR and ES forecasts

- Next 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:

$$\begin{aligned} Y_t|\mathcal{F}_{t-1} &\sim F_t \\ \left[VaR_t^{\dagger}, ES_t^{\dagger}\right] &\equiv \left[\quad F_t^{-1}(\alpha)\ ,\ \mathbb{E}\left[Y_t|Y_t \leq VaR_t^{\dagger}, \mathcal{F}_{t-1}\right] \quad\right] \end{aligned}$$

- ES is not elicitable, but can be elicited *jointly* with VaR (Fissler and Ziegel, 2016, *AoS*). We use the "FZ0" loss function for this purpose:

$$L_{FZ0}(y, v, e; \alpha) = -\frac{1}{\alpha e}\mathbf{1}\{y \leq v\}(v - y) + \frac{v}{e} + \log(-e) - 1$$

and so

$$\left[VaR_t^{\dagger}, ES_t^{\dagger}\right] = \arg\min_{(v,e)} \ \mathbb{E}\left[L_{FZ0}(Y_{t+1}, v, e; \alpha)|\mathcal{F}_t\right]$$

VaR and ES forecasts

- We consider a simple GARCH model as the baseline model for VaR and ES:

$$\begin{aligned} Y_t &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= \omega + \beta\sigma_{t-1}^2 + \alpha Y_{t-1}^2 \\ \text{and so} \quad [VaR_t, ES_t] &= \sigma_t \times [a, b] \\ \text{where} \quad a &= F_\varepsilon^{-1}(\alpha) \\ b &= \mathbb{E}[\varepsilon_t | \varepsilon_t \leq a] \end{aligned}$$

- We estimate these parameters by minimizing the FZ0 loss function
  - "Localizing" the (a, b) parameters works poorly (unsurprisingly)
  - We instead estimate these using the EDF of $\hat{\varepsilon}_t$
  - We localize only the GARCH parameters, keep $\left(\hat{a}, \hat{b}\right)$ fixed

VaR and ES forecasts

Benchmark non-local method is ranked 9th; Best local estimator beats with t-stat of -3.2

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

VaR and ES forecasts

Ex post best non-local method is ranked 4th, and is also in the MCS

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

Yield curve forecasts

- Finally we consider forecasts of the yield curve using the "dynamic Nelson-Siegel" model proposed by Diebold and Li (2006, *JoE*).
- The Nelson and Siegel (1987, *J.Bus*) model for a term structure of yields:

$$y_t(\tau) = \beta_{1t} + \beta_{2t}\left(\frac{1-e^{-\lambda_t\tau}}{\lambda_t\tau}\right) + \beta_{3t}\left(\frac{1-e^{-\lambda_t\tau}}{\lambda_t\tau} - e^{-\lambda_t\tau}\right) + u_t$$

- Parameters can be estimated from the yield curve each day.
  - When λ_t is fixed, $[\beta_{1t}, \beta_{2t}, \beta_{3t}]$ can be estimated by OLS, else by NLLS
- Diebold and Li (2006) propose estimating AR(1) processes for $\left\{\beta_{i,t}\right\}$

$$\beta_{\ i,t+1} = \phi_{0i} + \phi_{1i}\beta_{i,t} + \varepsilon_{i,t+1}$$

  - Inserting the predicted betas into the Nelson-Siegel model provides a foreacast of the next-period yield curve.

Yield curve forecasts

- We consider local versions of the DNS model, where the AR(1) models are estimated via local OLS.
    - We use the same state variable and same bandwidth for all three AR(1) models, although that could be relaxed.
- We consider US govt bonds with maturities of three and six months, and 1 to 10 years, a total of 12 maturities.
- We summarize the predictive performance of this model by summing the squared OOS forecast errors across maturities.
- We consider 1-day and 20-day forecast horizons

Yield curve forecasts, 1-day horizon

Best V-sample local estimator ranked fourth, beats benchmark with t-stat of -12.9. But...

	Method details		*Forecast performance*		
Rank	**StateVar**	**Window**	**AvgLoss**	**GW stat**	**MCS**
1	-	500	0.157	-9.499	✓
2	-	250	0.158	-5.071	✓
3	time,VIX	full	0.158	-11.618	×
4*	time, RV	full	0.158	-12.868	×
5	time,10Y-2Y	full	0.158	-14.128	×
6	time,FFR	full	0.158	-10.490	×
7	time	full	0.158	-14.523	×
8	RV	full	0.158	-9.099	×
9	-	1000	0.158	-1.605	×
10	FFR	full	0.158	-2.721	×
11	VIX	full	0.158	-4.440	×
12	10Y-2Y	full	0.158	-5.881	×
13	-	full	0.158	★	×

Yield curve forecasts, 1-day horizon

Best non-local estimator ranked *first*. Second-best is also non-local.

	Method details		*Forecast performance*		
Rank	**StateVar**	**Window**	**AvgLoss**	**GW stat**	**MCS**
1	-	500	0.157	-9.499	✓
2	-	250	0.158	-5.071	✓
3	time,VIX	full	0.158	-11.618	×
4*	time, RV	full	0.158	-12.868	×
5	time,10Y-2Y	full	0.158	-14.128	×
6	time,FFR	full	0.158	-10.490	×
7	time	full	0.158	-14.523	×
8	RV	full	0.158	-9.099	×
9	-	1000	0.158	-1.605	×
10	FFR	full	0.158	-2.721	×
11	VIX	full	0.158	-4.440	×
12	10Y-2Y	full	0.158	-5.881	×
13	-	full	0.158	★	×

Yield curve forecasts, 1-day horizon

- Local forecasts are beaten by the non-local forecasts. What's going on?
  - The baseline model is "too good" to be helped:

#	StateVar	Window	In-sample R^2
1	–	full	0.964

Yield curve forecasts, 1-day horizon

- Local forecasts are beaten by the non-local forecasts. What's going on?
  - The baseline model is "too good" to be helped:

#	StateVar	Window	In-sample R^2
1	–	full	0.964
2	–	1000	0.964
3	–	500	0.964
4	–	250	0.964
5	RV	full	0.964
6	VIX	full	0.964
7	FFR	full	0.964
8	10Y-2Y	full	0.964
9	time	full	0.964
10	time,RV	full	0.964
11	time,VIX	full	0.964
12	time,FFR	full	0.964
13	time,10Y-2Y	full	0.964

Yield curve forecasts, 20-day horizon

Benchmark non-local estimator ranked 8th; Best V-sample local estimator beats with t-stat of -6.9

	Method details		*Forecast performance*		
Rank	**StateVar**	**Window**	**AvgLoss**	**GW stat**	**MCS**
1	time	full	0.241	-6.542	✓
=1	time,FFR	full	0.241	-6.542	✓
=1	time,10Y-2Y	full	0.241	-6.542	✓
4	time,RV	full	0.242	-6.304	×
5*	time,VIX	full	0.244	-6.911	×
6	VIX	full	0.248	-2.399	×
7	10Y-2Y	full	0.250	-0.422	×
8	-	full	0.250	★	×
=8	FFR	full	0.250	0.000	×
10	RV	full	0.250	1.095	×
11	-	500	0.250	0.172	×
12	-	1000	0.253	1.364	×
13	-	250	0.262	2.567	×

Conditional forecast comparison

- All of the above rankings were done *unconditionally*, using average OOS loss.

- But if the forecaster has an idea for a state variable that is useful for tilting the model parameters, it might also be useful for predicting *which* model is likely to outperform in the next period.

- We now turn to *conditional* forecast comparisons, using:

  1. Giacomini-White (2005) parametric conditional comparions
  2. Li, Liao and Quaedvlieg (2021) nonparametric comparisons
  3. Nonparametric kernel smooths of OOS loss

- In all cases we compare the benchmark non-local model with the local model that performed best in the validation sample, and we use as the conditioning variable the state variable that is used in the local method.
  - Across all five comparisons this variable is either RV or VIX

Conditional forecast comparisons

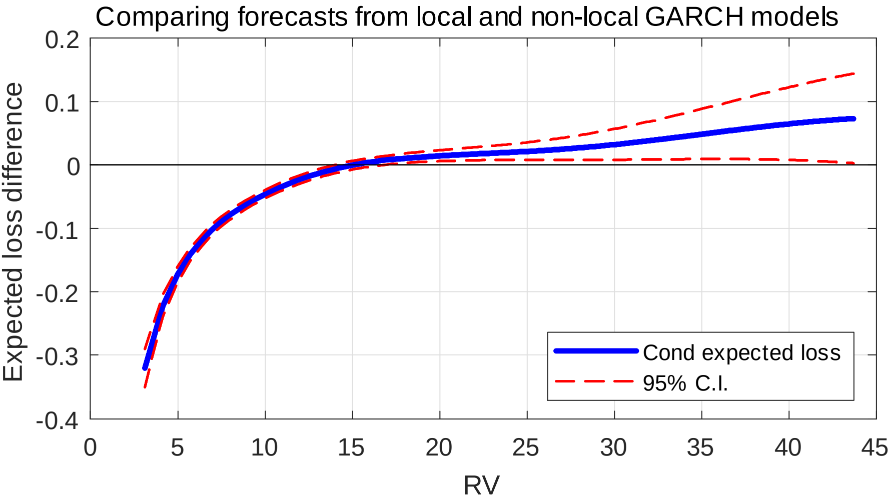
Positive slope implies local method relatively worse when volatility is high

$$L_{t+1}^{Local} - L_{t+1}^{Non-Local} = \beta_0 + \beta_1 \left(S_t - \bar{S}\right) + e_t$$

				Yield curve	
	GARCH	**HAR**	**VaR-ES**	**h=1**	**h=20**
$\hat{\beta}_0$	−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]
$\hat{\beta}_1$	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]

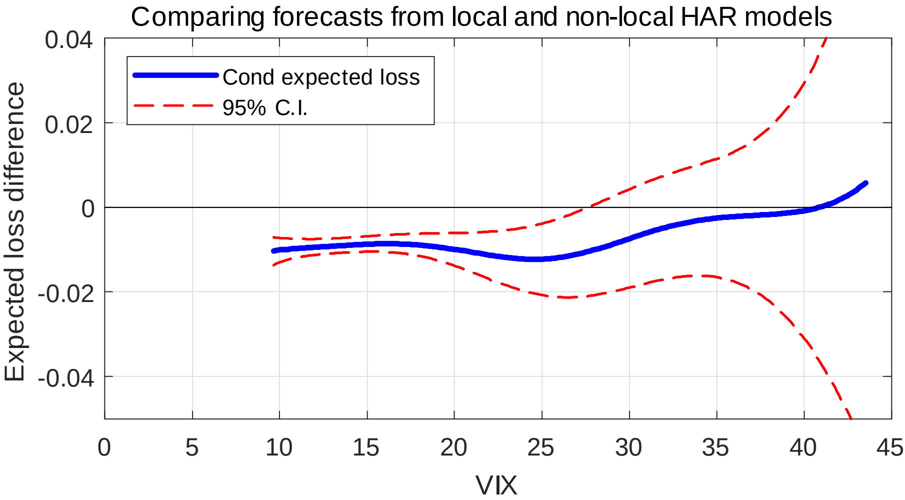
Conditional forecast comparisons: GARCH

Strong outperformance when RV is low, slight underperformance when RV is high



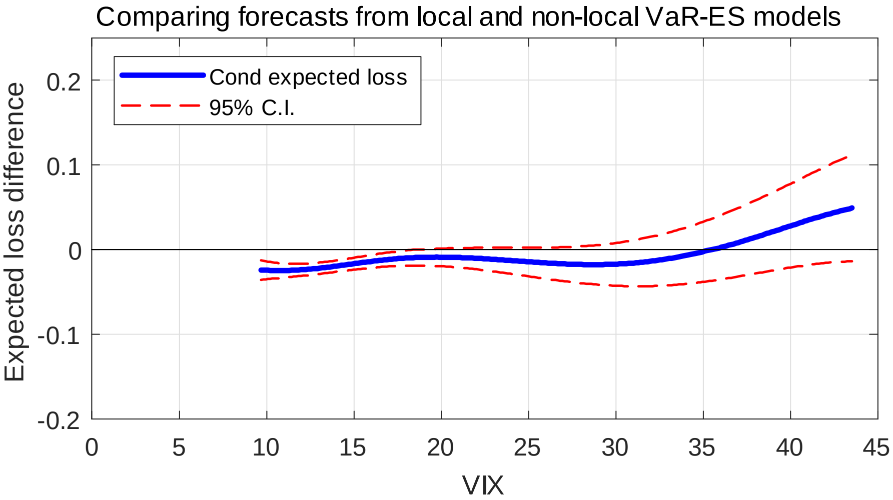
Conditional forecast comparisons: HAR

Strong outperformance when RV is low, approx equal performance when RV is high



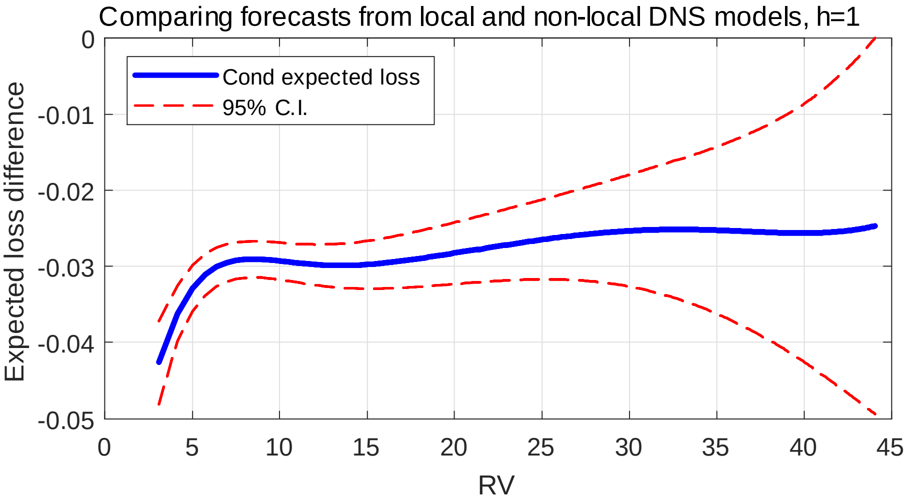
Conditional forecast comparisons: VaR and ES

Limited power in this application (5% tails are hard to learn about!)



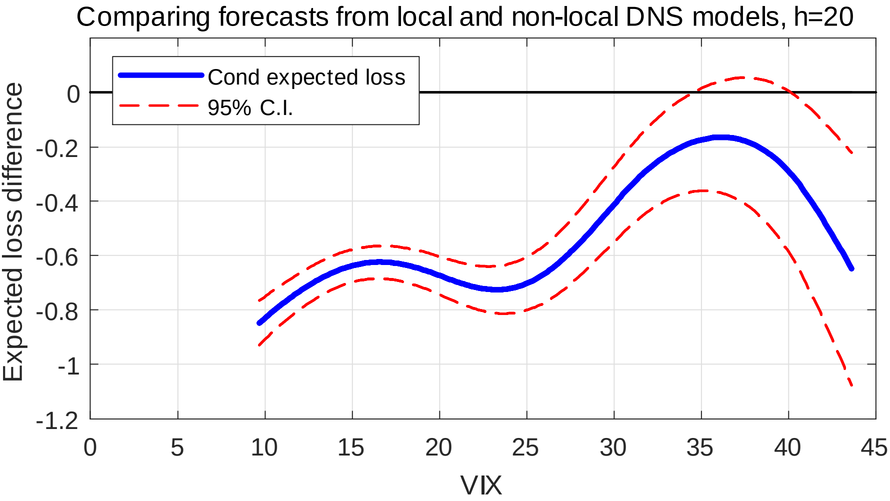
Conditional forecast comparisons: yield curve (h=1)

Local beats the benchmark across the entire support of RV



Conditional forecast comparisons: yield curve (h=1)

Local beats the benchmark across almost the entire support of VIX



Summary

- We suggest local estimation to improve the forecasts from a misspecified forecasting model, without altering the form of the model.
  - For various reasons, it may be hard to swap the model for a different one.
- We theoretically compare OOS forecasts from the local and standard estimation methods and observe a bias-variance trade-off.
  - Local methods are more likely to be helpful when the baseline model is not "too good," and when the state variable is not "too bad."
- We apply the proposed method to four economic forecasting problems, and find statistically significant improvements in *almost* all cases.
  - The level of vol is useful for risk forecasts, and also for yield curve forecasts.
  - Downweighting old observations is useful everywhere.

Appendix

GARCH-X forecasts

Best non-local method ranked 7th; Best V-sample local method beats with t-stat of -1.8

	Method details				*Forecast performance*		
Rank	**Model**	**StateVar**	**Window**		**AvgLoss**	**GW stat**	**MCS**
1	GARCH-X	time,RV	full		0.293	-9.329	✓
2	GARCH-X	RV	full		0.294	-9.382	✓
3	GARCH-X	time,FFR	full		0.309	-5.017	✓
4	GARCH-X	time,VIX	full		0.313	-4.013	✓
5	GARCH-X	time	full		0.313	-4.653	✓
6	GARCH	time,RV	full		0.320	-4.890	×
7	GARCH-X	-	250		0.324	-4.999	×
8	GARCH	RV	full		0.325	-4.239	×
9^*	GARCH-X	time,10Y-2Y	full		0.329	-1.828	×
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
17	GARCH-X	-	full		0.359	★	×
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
26	GARCH	-	full		0.402	4.844	×

HAR-X forecasts

Best non-local method ranked 7th; Best V-sample local method beats with t-stat of -6.8

	Method details				*Forecast performance*		
Rank	**Model**	**StateVar**	**Window**		**AvgLoss**	**GW stat**	**MCS**
1	HAR-X	RV	full		0.232	-7.001	✓
2*	HAR-X	time,RV	full		0.232	-6.843	✓
3	HAR-X	VIX	full		0.236	-6.722	×
4	HAR-X	time,10Y-2Y	full		0.241	-6.085	×
5	HAR-X	time,VIX	full		0.245	-5.723	×
6	HAR	time,VIX	full		0.246	-5.256	×
7	HAR-X	-	250		0.248	-5.576	×
8	HAR-X	time	full		0.248	-5.433	×
9	HAR	VIX	full		0.252	-4.829	×
10	HAR	time	full		0.252	-4.909	×
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
24	HAR-X	-	full		0.325	★	×
25	HAR-X	10Y-2Y	full		0.351	2.564	×
26	HAR-X	FFR	full		0.372	3.734	×