Selecting a Model for Forecasting

Jennifer L. Castle
with Jurgen A. Doornik and David F. Hendry

Magdalen College and Institute for New Economic Thinking at the Oxford Martin School, University of Oxford.

2nd Forecasting at Central Banks Conference
15-16 November 2018
Bank of England, London

Little agreement on 'best' models for real-world forecasting in wide-sense non-stationary settings facing shifts.

- Forcasting models used range from very parsimonious to large systems, machine learning and model or forecast averaging.

Many criteria proposed to select models with 'optimal' properties for forecasting in **stationary processes**, e.g. Akaike (1973).

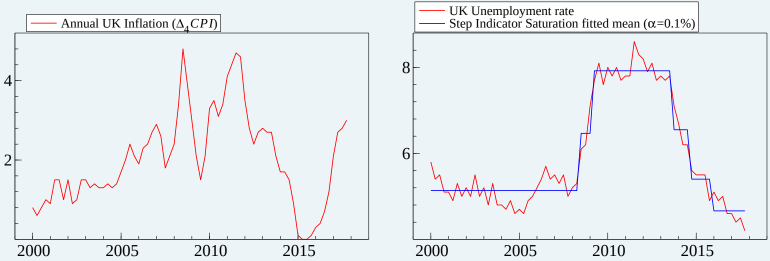
- **Yet even less agreement on selecting models in practice.**

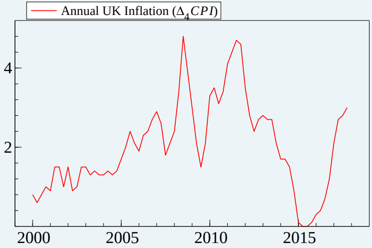
Explanation: distributional shifts differentially affect alternative formulations: (Clements and Hendry, 2001).

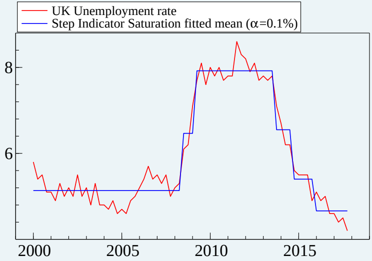
Contribution of this paper:

- In stationary static setting, strongly exogenous stochastic regressors, constant parameters implies retain regressors for forecasting if non-centralities $\psi > 1$.
  Does this trade-off hold if breaks?
- What is the 'optimal' nominal significance level α when **selecting** linear regression models for forecasting in data subject to breaks.

Generic trade-off between inconsistency and estimation uncertainty based on observed statistical significance.


Annual UK Inflation ($\Delta_4 CPI$)
4
2
2000
2005
2010
2015
UK Unemployment rate
Step Indicator Saturation fitted mean (α=0.1%)
8
6
2000
2005
2010
2015





What α minimises MSFE?

$$\mathsf{M}_1: \quad \pi_{t+1} - \pi_t = \mu + \beta_\pi \Delta\pi_t + \beta_{U_r} U_{r,t} + \nu_{1,t+1}$$

$$\mathsf{M}_2: \quad \pi_{t+1} - \pi_t = \mu + \gamma_\pi \Delta\pi_t + \nu_{2,t+1}$$

$\mathsf{H}_0 : \beta_{U_r} = 0$. Retain $U_{r,t}$ for forecasting if $\mathsf{t}^2_{\beta_{U_r}} > c^2_\alpha$.

Allow for breaks/outliers, and additional covariates: in practice add dynamics & non-linearities in non-congruent models.

1. **No breaks: forecasting with a stationary DGP**
2. Out-of-sample break – what is the impact of selection?
3. End-of-sample break – the impact of selection on different forecasting devices
4. Simulation evidence
5. Conclusions

(1) No breaks	
(2) Breaks:	
Out-of-sample (break at $T+1$)	End-of-sample (break at T)
(i) known regressors	
(ii) in-sample mean forecast	(ii) in-sample mean forecast
(iii) random walk forecast	(iii) random walk forecast

OXFORD MARTIN SCHOOL UNIVERSITY OF OXFORD Institute for New Economic Thinking AT THE OXFORD MARTIN SCHOOL

DGP given by VAR:

$$\begin{pmatrix} 1 & -\beta_1 & -\beta_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_t \\ x_{1,t} \\ x_{2,t} \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \epsilon_t \\ \eta_{1,t} \\ \eta_{2,t} \end{pmatrix}$$

where $\mathbf{y}_t = (y_t : x_{1,t} : x_{2,t})' \sim \mathsf{IN}_3 [\boldsymbol{\mu}, \boldsymbol{\Sigma}]$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_\epsilon & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}$

Let $\widehat{\mu}_i$ be sufficiently precise to neglect sampling variation so that $\mathsf{E}[y_t] = \mu_y = \beta_0 + \beta_1\mu_1 + \beta_2\mu_2$, and $\boldsymbol{\mu} = (\mu_y : \mu_1 : \mu_2)'$.

OXFORD MARTIN SCHOOL UNIVERSITY OF OXFORD Institute for New Economic Thinking AT THE OXFORD MARTIN SCHOOL

When to drop a regressor from the forecasting model?

$$\begin{aligned} \mathsf{M}_1 &: \quad y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \epsilon_t \\ \mathsf{M}_2 &: \quad y_t = \phi_0 + \gamma_1 x_{1,t} + \nu_t \end{aligned}$$

Choice between M_1 and M_2 depends on test of significance of $x_{2,t}$, where $\psi^2 = \frac{T\beta_2^2(1-\rho^2)}{\sigma_\epsilon^2}$ is squared population non-centrality of $\mathsf{t}_{\beta_2=0}$, under $\mathsf{H}_0 : \beta_2 = 0$.

When to drop a regressor from the forecasting model?

$$\begin{aligned} \mathsf{M}_1 &: \quad y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \epsilon_t \\ \mathsf{M}_2 &: \quad y_t = \phi_0 + \gamma_1 x_{1,t} + \nu_t \end{aligned}$$

Choice between M_1 and M_2 depends on test of significance of $x_{2,t}$, where $\psi^2 = \frac{T\beta_2^2(1-\rho^2)}{\sigma_\epsilon^2}$ is squared population non-centrality of $\mathsf{t}_{\beta_2=0}$, under $\mathsf{H}_0 : \beta_2 = 0$.

Compare 1-step ahead MSFE for known future regressors:

$$\mathsf{MSFE}_1 = \sigma_\epsilon^2 \left(1 + \frac{3}{T}\right) \quad \text{v} \quad \mathsf{MSFE}_2 = \sigma_\nu^2 \left(1 + \frac{2}{T}\right)$$

where $\sigma_\nu^2 = \sigma_\epsilon^2 \left(1 + T^{-1}\psi^2\right) \geq \sigma_\epsilon^2$ and $\sigma_\nu^2 \to \sigma_\epsilon^2$ as T increases for a given ψ.

When to drop a regressor from the forecasting model?

$$\mathsf{M}_1 \quad : \quad y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \epsilon_t$$
$$\mathsf{M}_2 \quad : \quad y_t = \phi_0 + \gamma_1 x_{1,t} + \nu_t$$

Choice between M_1 and M_2 depends on test of significance of $x_{2,t}$, where $\psi^2 = \frac{T\beta_2^2\left(1-\rho^2\right)}{\sigma_\epsilon^2}$ is squared population non-centrality of $\mathsf{t}_{\beta_2=0}$, under $\mathsf{H}_0 : \beta_2 = 0$.

Compare 1-step ahead MSFE for known future regressors:

$$\mathsf{MSFE}_1 = \sigma_\epsilon^2\left(1+\frac{3}{T}\right) \quad \text{v} \quad \mathsf{MSFE}_2 = \sigma_\nu^2\left(1+\frac{2}{T}\right)$$

where $\sigma_\nu^2 = \sigma_\epsilon^2\left(1+T^{-1}\psi^2\right) \geq \sigma_\epsilon^2$ and $\sigma_\nu^2 \to \sigma_\epsilon^2$ as T increases for a given ψ.

M_2 has one fewer parameter to estimate, traded off against a larger equation variance.

When does parsimony pay in forecasting?

For $\mathsf{MSFE}_2 \leq \mathsf{MSFE}_1$ requires:

$$\sigma_\nu^2\left(1+\frac{2}{T}\right)-\sigma_\epsilon^2\left(1+\frac{3}{T}\right)=\frac{\sigma_\epsilon^2}{T}\left[\psi^2\left(1+\frac{2}{T}\right)-1\right]\leq 0$$

which occurs when $\psi^2 \leq T/(T+2)$ (independent of ρ).

If $\psi > 1$, information content of $x_{2,t}$ outweighs parameter estimation cost for 1-step forecasts, regardless of $|\rho| < 1$ between x_1 and x_2.

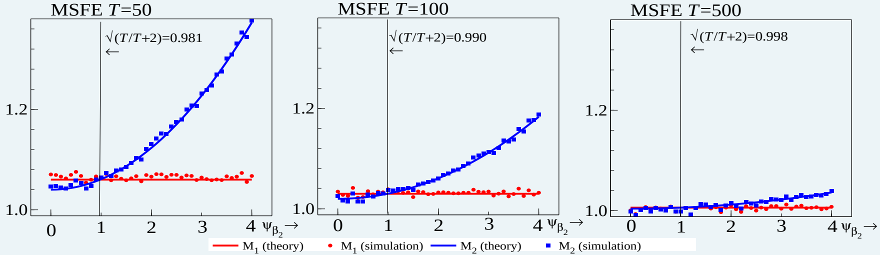
When does parsimony pay in forecasting?

For $\mathsf{MSFE}_2 \leq \mathsf{MSFE}_1$ requires:

$$\sigma_\nu^2 \left(1 + \frac{2}{T}\right) - \sigma_\epsilon^2 \left(1 + \frac{3}{T}\right) = \frac{\sigma_\epsilon^2}{T}\left[\psi^2 \left(1 + \frac{2}{T}\right) - 1\right] \leq 0$$

which occurs when $\psi^2 \leq T/(T+2)$ (independent of ρ).

If $\psi > 1$, information content of $x_{2,t}$ outweighs parameter estimation cost for 1-step forecasts, regardless of $|\rho| < 1$ between x_1 and x_2.



But DGP never known, so in practice need to select between M_1 & M_2.

Forecasts from selected model, called M_3, based on a mixture of M_1 and M_2 in repeated sampling depending on ψ^2 and α.

MSFE for model selection

$$\begin{aligned}\mathsf{MSFE}_3 &= \mathsf{p}_\alpha[\psi]\mathsf{MSFE}_1 + (1-\mathsf{p}_\alpha[\psi])\,\mathsf{MSFE}_2 \\ &= \mathsf{MSFE}_1 + (1-\mathsf{p}_\alpha[\psi])\,(\mathsf{MSFE}_2 - \mathsf{MSFE}_1)\end{aligned}$$

where $\mathsf{p}_\alpha[\psi] = \Pr\left(\mathsf{t}^2_{\beta_2=0} \geq c^2_\alpha\right)$

Forecasts from selected model, called M_3, based on a mixture of M_1 and M_2 in repeated sampling depending on ψ^2 and α.

MSFE for model selection

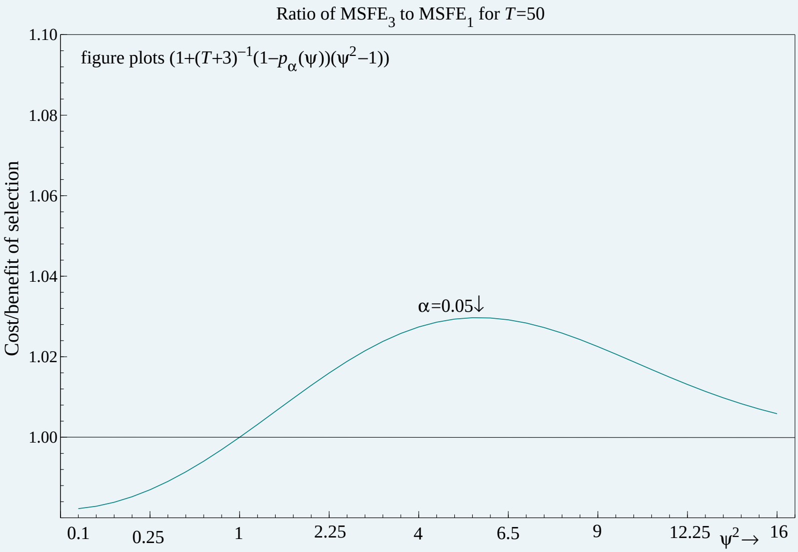
$$\begin{aligned}\mathsf{MSFE}_3 &= \mathsf{p}_\alpha[\psi]\mathsf{MSFE}_1 + (1 - \mathsf{p}_\alpha[\psi])\,\mathsf{MSFE}_2 \\ &= \mathsf{MSFE}_1 + (1 - \mathsf{p}_\alpha[\psi])\,(\mathsf{MSFE}_2 - \mathsf{MSFE}_1)\end{aligned}$$

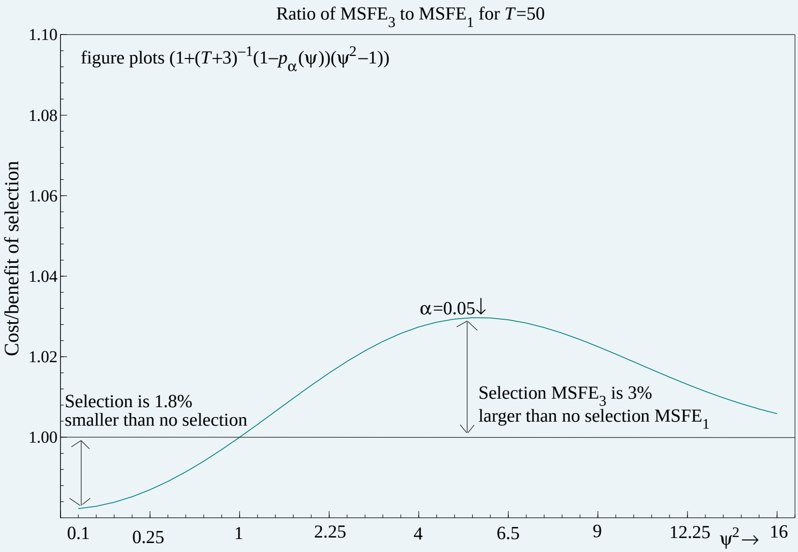
where $\mathsf{p}_\alpha[\psi] = \Pr\left(\mathsf{t}^2_{\beta_2=0} \geq c^2_\alpha\right)$

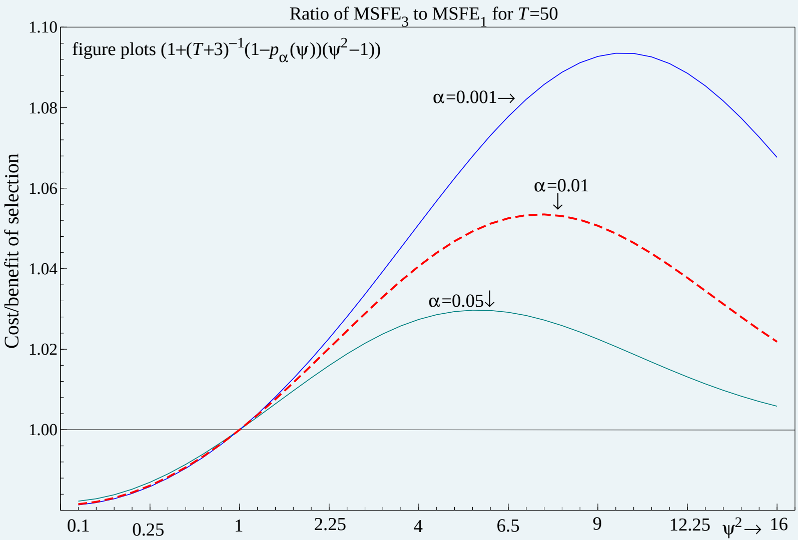
$$\boxed{\mathsf{MSFE}_3 \approx \mathsf{MSFE}_1 + \sigma^2_\epsilon T^{-1}\,(1 - \mathsf{p}_\alpha[\psi])\,(\psi^2 - 1)}$$

- $\mathsf{MSFE}_3 \leq \mathsf{MSFE}_1$ whenever $\psi^2 \leq 1$.
- MSFE_3 highly non-linear function of ψ^2 and α.

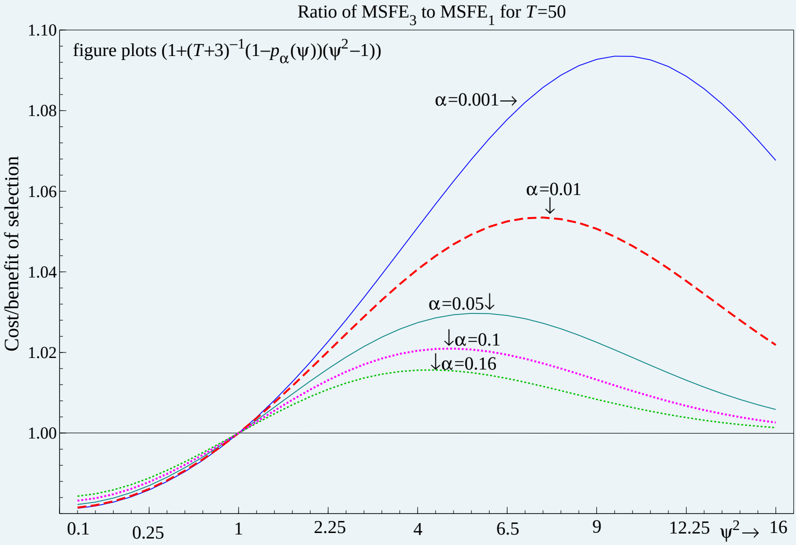
Costs/benefits of selection



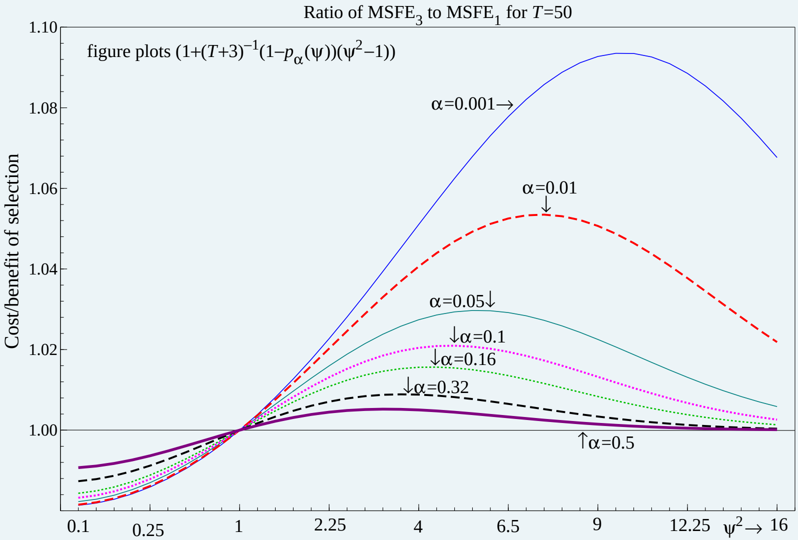

Ratio of MSFE$_3$ to MSFE$_1$ for T=50
figure plots $(1+(T+3)^{-1}(1-p_\alpha(\psi))(\psi^2-1))$
Cost/benefit of selection
1.10
1.08
1.06
1.04
1.02
1.00
α=0.05↓
Selection is 1.8%
smaller than no selection
Selection MSFE$_3$ is 3%
larger than no selection MSFE$_1$
0.1
0.25
1
2.25
4
6.5
9
12.25
$\psi^2 \rightarrow$
16


Ratio of MSFE3 to MSFE1 for T=50
figure plots (1+(T+3)−1(1−pα(ψ))(ψ2−1))
α=0.001→
α=0.01
α=0.05↓
Cost/benefit of selection
1.10
1.08
1.06
1.04
1.02
1.00
0.1
0.25
1
2.25
4
6.5
9
12.25
ψ2→
16

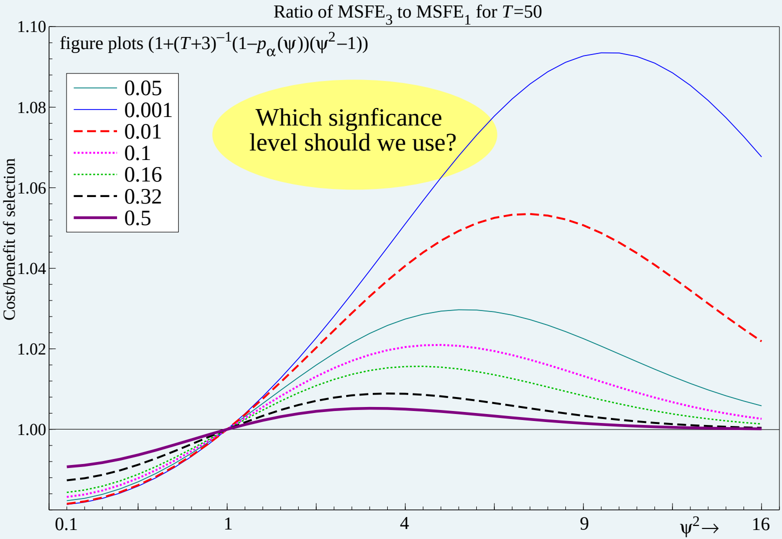
Institute for New Economic Thinking at the Oxford Martin School

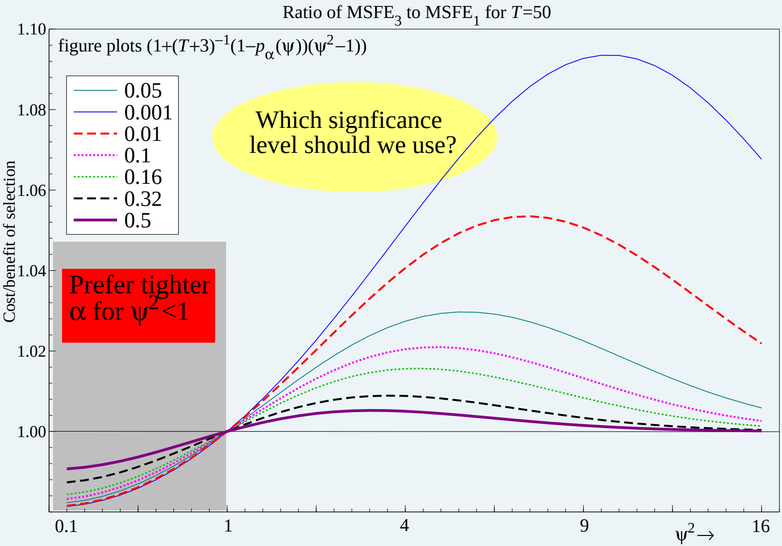


Costs/benefits of selection

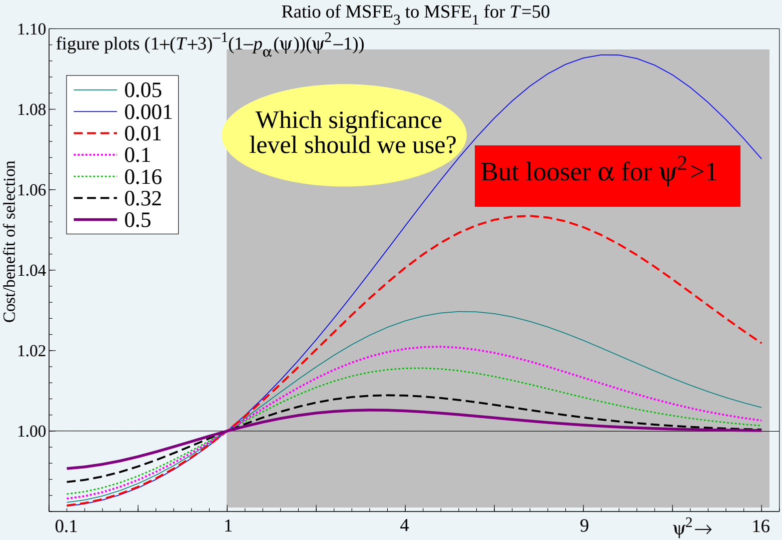


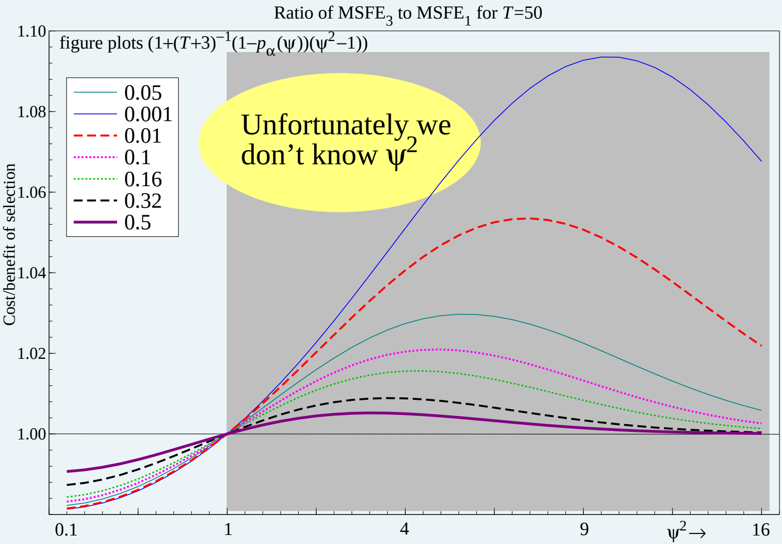
Costs/benefits of selection





Ratio of MSFE3 to MSFE1 for T=50
figure plots (1+(T+3)^−1(1−pα(ψ))(ψ^2−1))
0.05
0.001
0.01
0.1
0.16
0.32
0.5
Which signficance level should we use?
Prefer tighter α for ψ^2<1
Cost/benefit of selection
1.10
1.08
1.06
1.04
1.02
1.00
0.1
1
4
9
16
ψ^2→


Ratio of MSFE3 to MSFE1 for T=50
figure plots (1+(T+3)^−1(1−pα(ψ))(ψ^2−1))
0.05
0.001
0.01
0.1
0.16
0.32
0.5
Which signficance level should we use?
But looser α for ψ^2>1
Cost/benefit of selection
1.10
1.08
1.06
1.04
1.02
1.00
0.1
1
4
9
ψ^2→
16



Ratio of MSFE$_3$ to MSFE$_1$ for T=50
figure plots $(1+(T+3)^{-1}(1-p_\alpha(\psi))(\psi^2-1))$
0.05
0.001
0.01
0.1
0.16
0.32
0.5
Unfortunately we don't know ψ^2
Cost/benefit of selection
1.10
1.08
1.06
1.04
1.02
1.00
0.1
1
4
9
$\psi^2\rightarrow$
16

Trade-off: tighter α lowers MSFE for $\psi^2 < 1$ by eliminating x_2 more frequently; looser α preferred for $\psi^2 > 1$ as x_2 more likely retained.

Two inequalities:

- $x_{2,t}$ omitted if $\mathsf{t}^2_{\beta_2=0} < c^2_\alpha$, which occurs when $\widehat{\beta}^2_2 < \frac{c^2_\alpha \sigma^2_\epsilon}{T(1-\rho^2)}$.
- $x_{2,t}$ omitted if $\psi^2 < 1$ for smaller MSFE.

Trade-off: tighter α lowers MSFE for $\psi^2 < 1$ by eliminating x_2 more frequently; looser α preferred for $\psi^2 > 1$ as x_2 more likely retained.

Two inequalities:

- $x_{2,t}$ omitted if $\mathsf{t}^2_{\beta_2=0} < c^2_\alpha$, which occurs when $\widehat{\beta}^2_2 < \frac{c^2_\alpha \sigma^2_\epsilon}{T(1-\rho^2)}$.
- $x_{2,t}$ omitted if $\psi^2 < 1$ for smaller MSFE.

Equating the two inequalities $\implies c^2_\alpha \leq 2$ and:

$$\boxed{\mathsf{E}[\mathsf{t}^2_{\beta_2=0}] = 2 \implies \alpha = 0.16}$$

AIC: LR χ^2 test, 2 nested models, 1df, penalty=2, $\rightarrow \alpha = 16\%$.

Trade-off: tighter α lowers MSFE for $\psi^2 < 1$ by eliminating x_2 more frequently; looser α preferred for $\psi^2 > 1$ as x_2 more likely retained.

Two inequalities:

- $x_{2,t}$ omitted if $\mathsf{t}^2_{\beta_2=0} < c^2_\alpha$, which occurs when $\widehat{\beta}^2_2 < \frac{c^2_\alpha \sigma^2_\epsilon}{T(1-\rho^2)}$.
- $x_{2,t}$ omitted if $\psi^2 < 1$ for smaller MSFE.

Equating the two inequalities $\implies$ $c^2_\alpha \leq 2$ and:

$$\boxed{\mathsf{E}[\mathsf{t}^2_{\beta_2=0}] = 2 \implies \alpha = 0.16}$$

AIC: LR χ^2 test, 2 nested models, 1df, penalty=2, $\rightarrow \alpha = 16\%$.

Results close to implied significance level for AIC in Campos, Hendry, and Krolzig (2003), Pötscher (1991), Leeb and Pötscher (2009).
Will also increase adventitious retention of irrelevant variables.
Trade-off dependent on how many likely to be relevant/irrelevant.

1. No breaks: forecasting with a stationary DGP
2. **Out-of-sample break – what is the impact of selection?**
3. End-of-sample break – the impact of selection on different forecasting devices
4. Simulation evidence
5. Conclusions

(1) No breaks	
(2) Breaks:	
Out-of-sample (break at $T+1$)	End-of-sample (break at T)
(i) known regressors	
(ii) in-sample mean forecast	(ii) in-sample mean forecast
(iii) random walk forecast	(iii) random walk forecast

Location shift out-of-sample

Location shift in x_2 at $T+1$ with the forecast origin of T:

$$x_{1,t} = \mu_1 + \eta_{1,t} \qquad t = 1, \ldots, T+1.$$

$$x_{2,t} = \begin{cases} \mu_2 + \eta_{2,t} & t = 1, \ldots, T \\ \mu_2 + \delta + \eta_{2,t} & t = T+1 \end{cases}$$

Location shift in x_2 at $T+1$ with the forecast origin of T:

$$x_{1,t} = \mu_1 + \eta_{1,t} \qquad t = 1, \ldots, T+1.$$

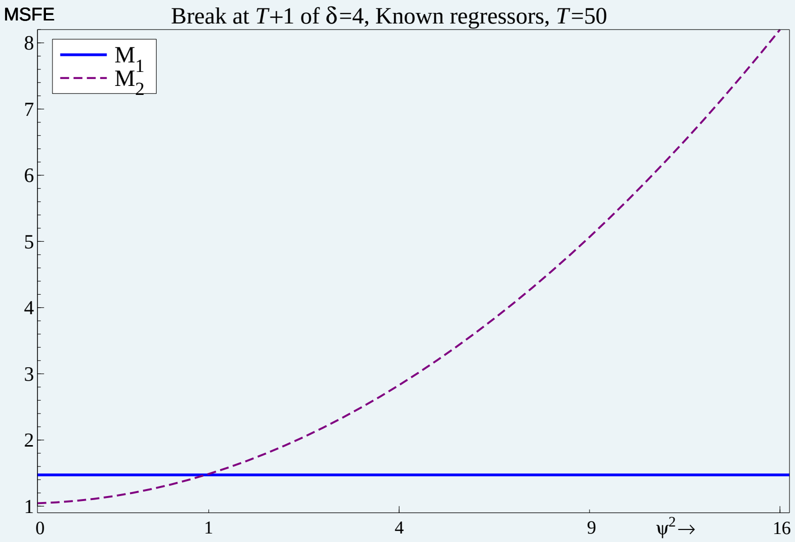
$$x_{2,t} = \begin{cases} \mu_2 + \eta_{2,t} & t = 1, \ldots, T \\ \mu_2 + \delta + \eta_{2,t} & t = T+1 \end{cases}$$

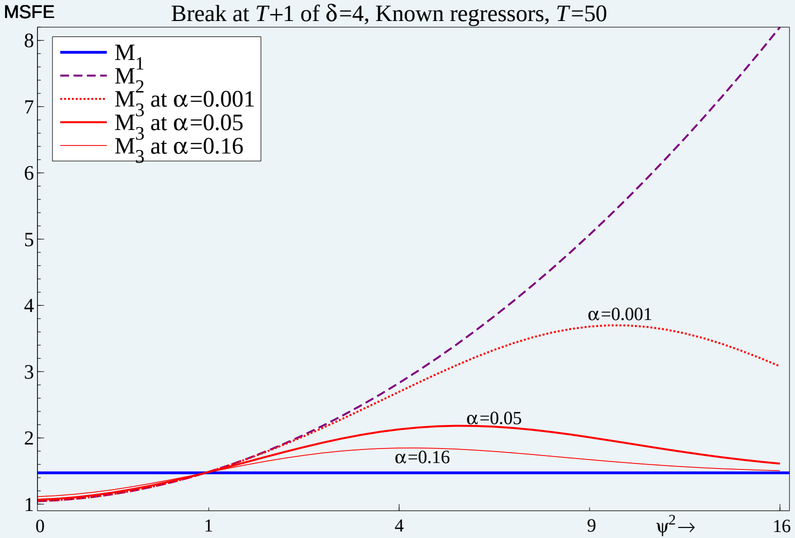
- **Known future values of regressors**

Break in μ_2 does not affect choice of forecasting model as break is captured in $x_{2,T+1}$.

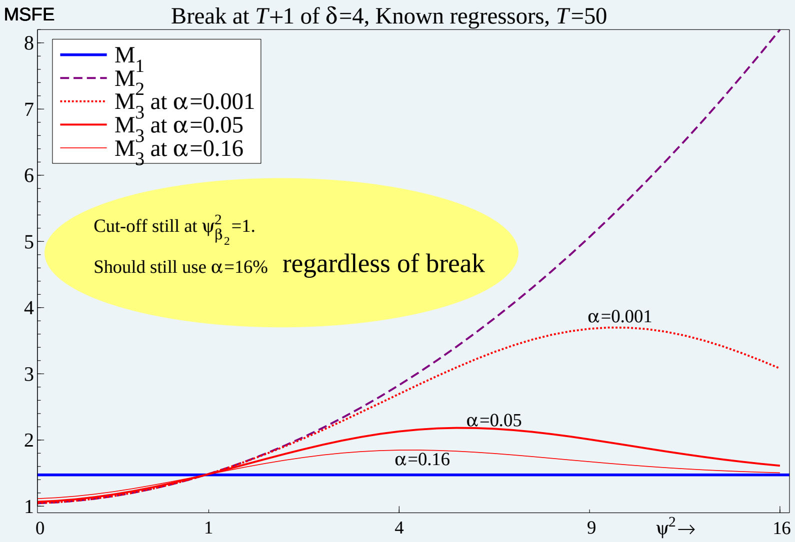
Trade-off at $\psi = 1$ holds regardless of break:

always (**never**) include for $\psi^2 \geq 1$ ($\psi^2 < 1$).


MSFE
Break at T+1 of δ=4, Known regressors, T=50
M1
M2
8
7
6
5
4
3
2
1
0
1
4
9
ψ2→
16


MSFE
Break at T+1 of δ=4, Known regressors, T=50
M1
M2
M3 at α=0.001
M3 at α=0.05
M3 at α=0.16
α=0.001
α=0.05
α=0.16
8
7
6
5
4
3
2
1
0
1
4
9
ψ2→
16

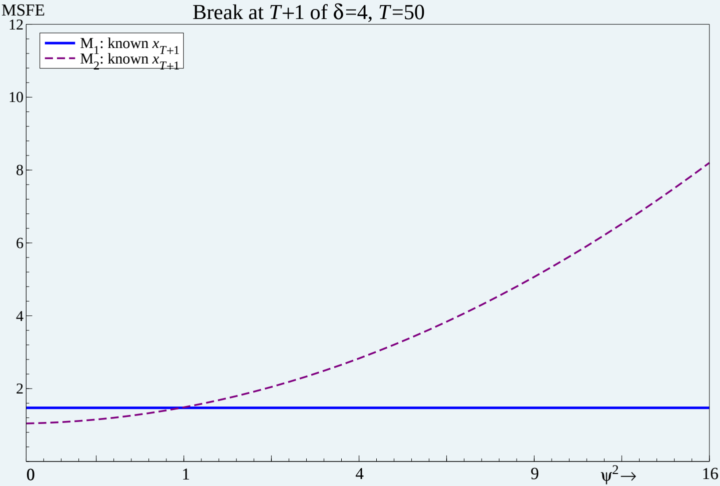
OXFORD MARTIN SCHOOL UNIVERSITY OF OXFORD Institute for New Economic Thinking AT THE OXFORD MARTIN SCHOOL



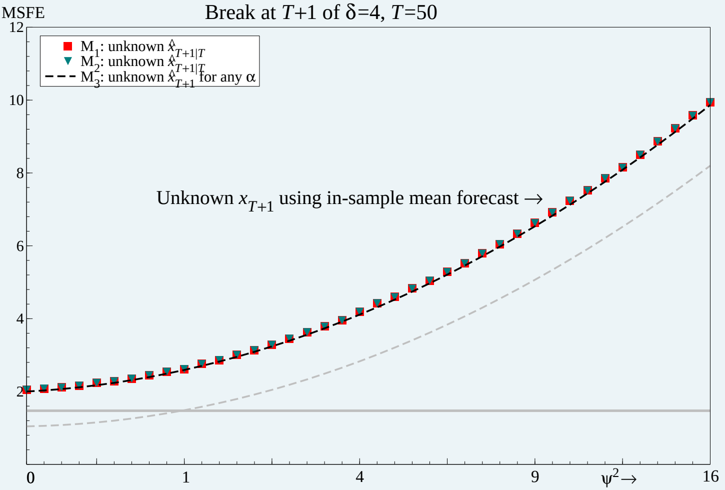
- **Unknown future values of regressors**

Link between y and x_i stays constant, but shift at $T+1$ not anticipated, inducing shift in $y_{T+1} \implies$ **forecast failure**.

In-sample mean forecast: μ_y shifts to $(\mu_y + \beta_2\delta)$ at $T+1$, but forecast to be μ_y.


Break at T+1 of δ=4, T=50
MSFE
12
10
8
6
4
2
M_1: known x_{T+1}
M_2: known x_{T+1}
0
1
4
9
$\psi^2 \rightarrow$
16

Unknown x_{t+1}, in-sample mean forecast



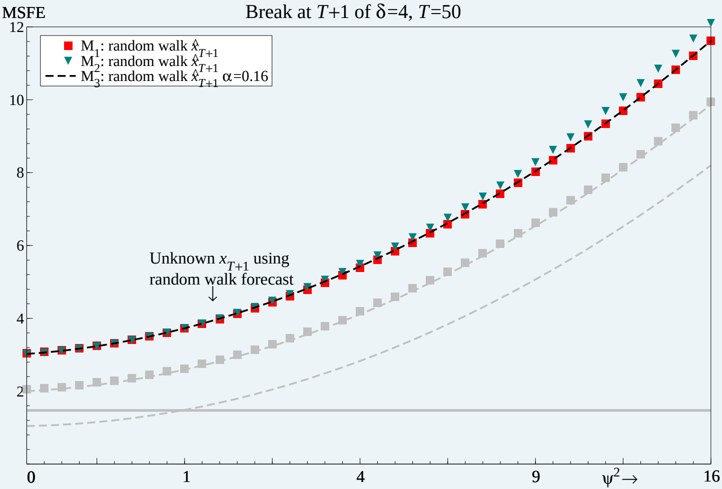
MSFE trajectories very similar: unanticipated break dominates any forecast error resulting from model mis-specification.

Selection has little effect. **Parsimony neither helps nor hinders**.

- **Unknown future values of regressors using random walk forecast**

Forecasts for exogenous variables: $\overline{x}_{i,T+1|T} = x_{i,T}, \; i = 1, 2$.

Last in-sample observation imprecise measure of out-of-sample mean, but unbiased when **no location shifts** (with no dynamics).



MSFE$_1$ and MSFE$_2$ very similar using random walk forecasts.

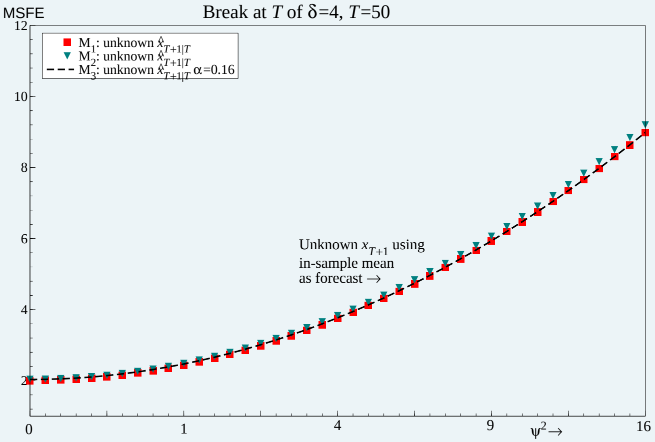
Worse than in-sample mean as both $\widehat{x}_{1,T+1}$ and $\widehat{x}_{2,T+1}$ incur cost.

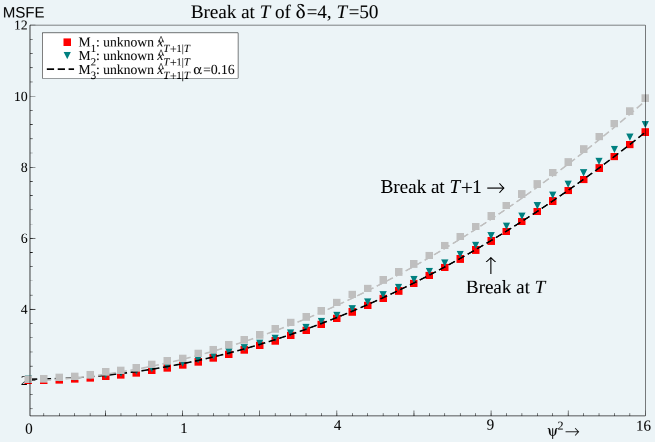
For selection, trade-off as before but switch point can be smaller than $\psi^2 = 1$, depending on the values of ρ and T – but impact very small.

1. No breaks: forecasting with a stationary DGP
2. Out-of-sample break – what is the impact of selection?
3. **End-of-sample break – the impact of selection on different forecasting devices**
4. Simulation evidence
5. Conclusions

<table>
<tr><td colspan="2">(1) No breaks</td></tr>
<tr><td colspan="2">(2) Breaks:</td></tr>
<tr><td>Out-of-sample (break at $T+1$)</td><td>End-of-sample (break at T)</td></tr>
<tr><td>(i) known regressors</td><td></td></tr>
<tr><td>(ii) in-sample mean forecast</td><td>(ii) in-sample mean forecast</td></tr>
<tr><td>(iii) random walk forecast</td><td>(iii) random walk forecast</td></tr>
</table>

Location shift in x_2 at T with the forecast origin of T





Similar to out-of-sample break.

Impact of break on estimated mean of $x_{2,t}$ small unless δ very large.
Cost of omitting x_2 rises with $(\beta_2\delta)^2$, but increased ψ^2 increases probability of retaining x_2, unconnected with magnitude of δ.

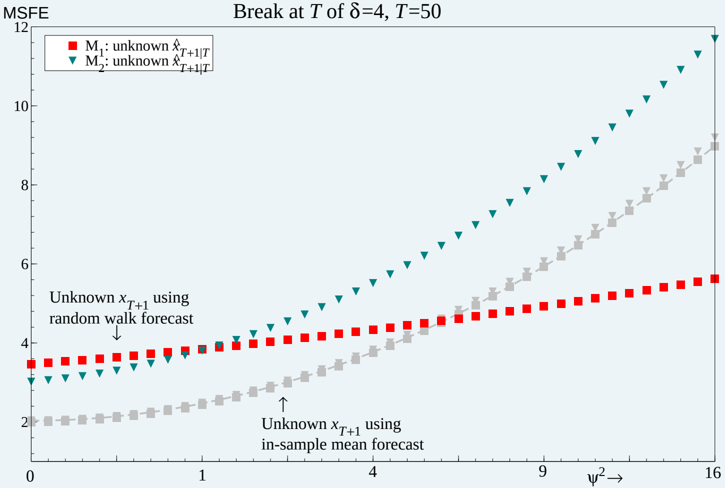
- **Unknown future values of regressors using random walk**

Random walk is now a **'robust forecasting device'**: improved forecasting properties following location shift.

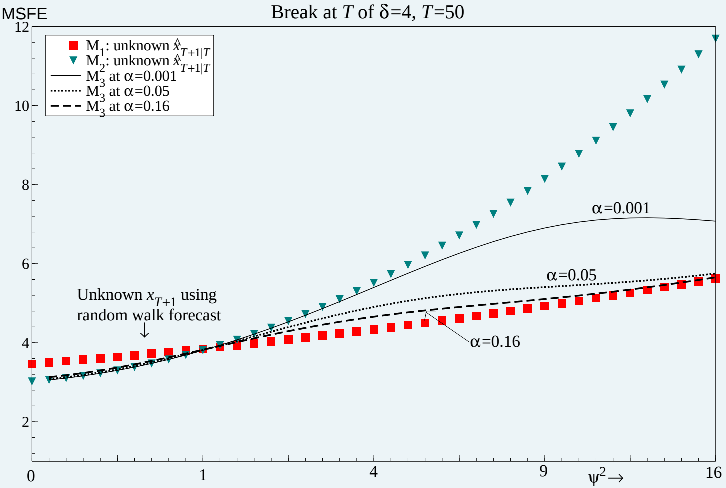
Forecasts for exogenous variables: $\overline{x}_{i,T+1|T} = x_{i,T},\ i = 1, 2$.

$\mathsf{E}\left[x_{1,T}\right] = \mu_1$, $\mathsf{E}\left[x_{2,T}\right] = \mu_2 + \delta$; $\mathsf{E}\left[\Delta x_{1,T+1}\right] = 0$, $\mathsf{E}\left[\Delta x_{2,T+1}\right] = 0$.

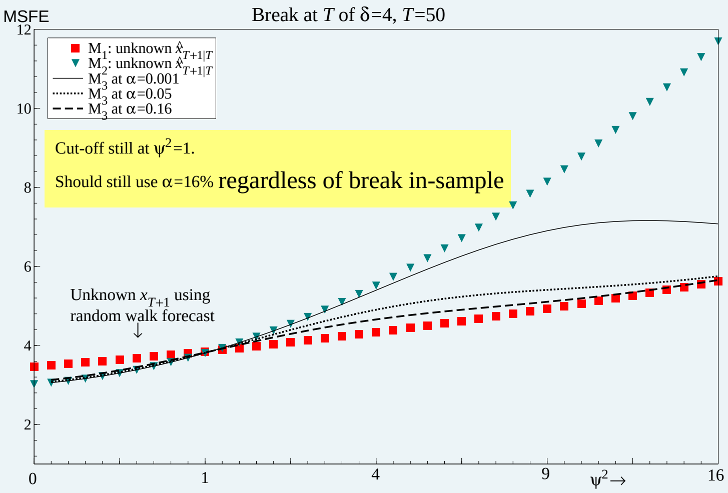
Unbiased forecasts for both $x_{1,T+1}$ and $x_{2,T+1}$ but inefficient forecast for $x_{1,T+1}$ relative to in-sample mean forecast as no shift.



Extra cost relative to mean forecast for small ψ^2 as robust $\hat{x}_{1,T+1}$ not needed – if known which regressors subject to break could improve.



Selection can be beneficial – retains relevant/eliminates irrelevant regressors that shift. Close to M_1 at $\alpha = 0.16$ for $\psi^2 > 1$ and to M_2 for $\psi^2 < 1$.



Selection can be beneficial – retains relevant/eliminates irrelevant regressors that shift. Close to M_1 at $\alpha = 0.16$ for $\psi^2 > 1$.

Summary of analytic results

$$\sigma_\epsilon^2 = 1,\ \beta_0 = 5,\ \beta_1 = 1,\ \mu_1 = \mu_2 = 2,\ \delta = 4,\ \rho = 0.5$$

Case		$\psi^2_{\beta_2} = 0$	$\psi^2_{\beta_2} = 1$	$\psi^2_{\beta_2} = 4$	$\psi^2_{\beta_2} = 16$
Stationary	M_2		1.001		
($\delta = 0$)	M_3		1.000		
Out of sample shift					
Known future regressors	M_2		1.014		
	M_3		1.009		
Unknown future regressors					
mean forecast	M_2		1.000		
	M_3		1.000		
random walk forecast	M_2		1.004		
	M_3		1.002		
In-sample shift					
mean forecast	M_2		**1.021**		
	M_3		1.014		
random walk forecast	M_2		**0.990**		
	M_3		0.993		

Figures reported are $\frac{\text{MSFE}_2}{\text{MSFE}_1}$ and $\frac{\text{MSFE}_3}{\text{MSFE}_1}$ for $T = 50$ and $\alpha = 0.16$.

- **Supports $\psi = 1$ as cut-off. Ratios very close to 1.**

$\sigma_\epsilon^2 = 1,\ \beta_0 = 5,\ \beta_1 = 1,\ \mu_1 = \mu_2 = 2,\ \delta = 4,\ \rho = 0.5$

Case		$\psi_{\beta_2}^2 = 0$	$\psi_{\beta_2}^2 = 1$	$\psi_{\beta_2}^2 = 4$	$\psi_{\beta_2}^2 = 16$
Stationary	M$_2$	0.981	1.001		
($\delta = 0$)	M$_3$	0.984	1.000		
Out of sample shift					
Known future regressors	M$_2$	**0.709**	1.014		
	M$_3$	0.756	1.009		
Unknown future regressors					
mean forecast	M$_2$	1.000	1.000		
	M$_3$	1.000	1.000		
random walk forecast	M$_2$	0.993	1.004		
	M$_3$	0.994	1.002		
In-sample shift					
mean forecast	M$_2$	**1.020**	**1.021**		
	M$_3$	1.017	1.014		
random walk forecast	M$_2$	0.871	**0.990**		
	M$_3$	0.892	0.993		

Figures reported are $\frac{\text{MSFE}_2}{\text{MSFE}_1}$ and $\frac{\text{MSFE}_3}{\text{MSFE}_1}$ for $T = 50$ and $\alpha = 0.16$.

- **M$_2$ correct model, but selection not costly.**
- **In some cases gains over M$_1$ very large.**

$\sigma_\epsilon^2 = 1,\ \beta_0 = 5,\ \beta_1 = 1,\ \mu_1 = \mu_2 = 2,\ \delta = 4,\ \rho = 0.5$

Case		$\psi_{\beta_2}^2 = 0$	$\psi_{\beta_2}^2 = 1$	$\psi_{\beta_2}^2 = 4$	$\psi_{\beta_2}^2 = 16$
Stationary	M_2	0.981	1.001	1.060	1.295
($\delta = 0$)	M_3	0.984	1.000	**1.016**	**1.001**
Out of sample shift					
Known future regressors	M_2	**0.709**	1.014	**1.927**	**5.582**
	M_3	0.756	1.009	**1.256**	**1.022**
Unknown future regressors					
mean forecast	M_2	1.000	1.000	1.000	1.000
	M_3	1.000	1.000	**1.000**	**1.000**
random walk forecast	M_2	0.993	1.004	1.020	1.043
	M_3	0.994	1.002	**1.006**	**1.000**
In-sample shift					
mean forecast	M_2	**1.020**	**1.021**	1.022	1.024
	M_3	1.017	1.014	**1.006**	**1.000**
random walk forecast	M_2	0.871	**0.990**	1.273	2.078
	M_3	0.892	0.993	**1.075**	**1.005**

Figures reported are $\frac{\text{MSFE}_2}{\text{MSFE}_1}$ and $\frac{\text{MSFE}_3}{\text{MSFE}_1}$ for $T = 50$ and $\alpha = 0.16$.

- **Costs of selection are usually small, irrespective of ψ.**
- **Model selection reduces risk relative to worst model.**

1. No breaks: forecasting with a stationary DGP
2. Out-of-sample break – what is the impact of selection?
3. End-of-sample break – the impact of selection on different forecasting devices
4. **Simulation evidence**
5. Conclusions

(1) No breaks	
(2) Breaks:	
Out-of-sample (break at $T+1$)	End-of-sample (break at T)
(i) known regressors	
(ii) in-sample mean forecast	(ii) in-sample mean forecast
(iii) random walk forecast	(iii) random walk forecast

Large simulation study looking across:

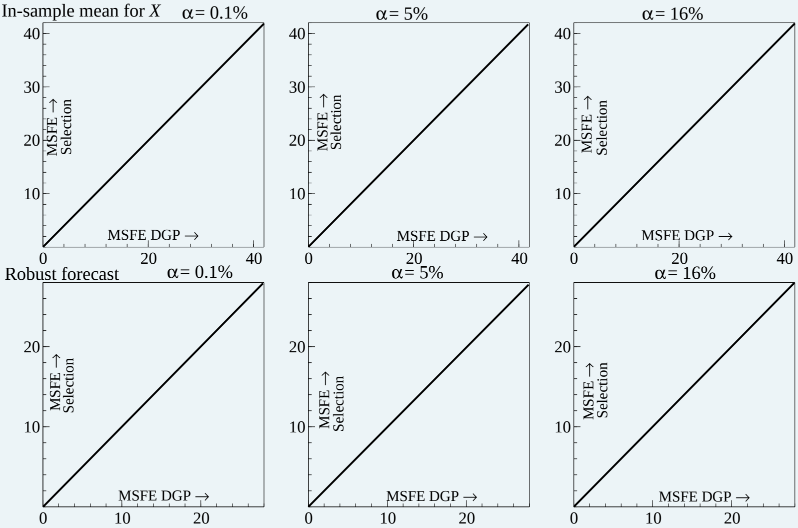
- Varying non-centralities and DGP sizes
- Varying numbers of relevant/irrelevant regressors
- Varying sample size & break magnitude
- Breaks in relevant/irrelevant/all regressors
- Breaks in mean/persistence
- Breaks in/out-of sample
- Range of forecasting models including in-sample mean & robust

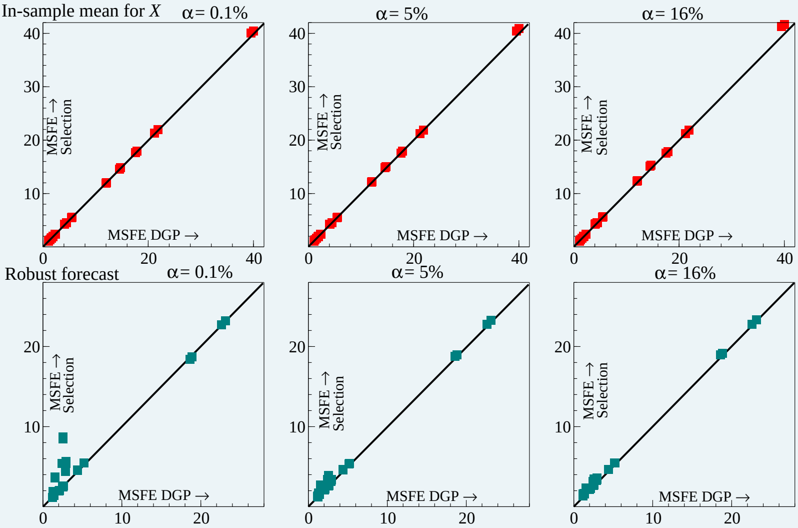
Selection by *Autometrics* for $\alpha = (0.001, 0.01, 0.05, 0.1, 0.16, 0.32, 0.5)$

Results:

2142 distinct MSFE observations, $\overline{\text{MSFE}} = 5.15$ and $\sigma = 7.50$.

Is selection costly?





Knowing DGP infeasible – selection must be undertaken.

Incurs almost no cost relative to DGP if α not too tight.

Theory: retain if $\psi > 1 \implies \mathrm{t}^2 > 2$, regardless of location shifts.

Looser than conventional significance levels:

fewer relevant variables excluded contributing to forecast accuracy

more irrelevant variables retained by chance, but coefficient estimates driven towards zero when updating

Theory: retain if $\psi > 1 \implies \mathrm{t}^2 > 2$, regardless of location shifts.

Looser than conventional significance levels:

fewer relevant variables excluded contributing to forecast accuracy
more irrelevant variables retained by chance, but coefficient estimates driven towards zero when updating

Simulation evidence provides guidance for forecasting

- Support for selecting models $\approx 10\%$, $N = 15$ or 16% at $N = 2$.
- Knowing DGP but forecasting $\mathbf{x}$ rarely delivered best MSFE.
- In-sample mean for $\mathbf{x}$: worst model for end-of-sample breaks in relevant/all regressors, but best out-of-sample.
- RW with difference robust forecast for $\mathbf{x}$: best for end-of-sample breaks in relevant/all regressors, poor if breaks out-of-sample.
- Direct AR(1) forecast for y: best if breaks in irrelevant variables.
- Simulation highlighted complexity of selection rule for forecasting – highly non-linear with many interaction terms. Results depended on all aspects of experimental design especially retention probability given ψ.

OXFORD MARTIN SCHOOL | UNIVERSITY OF OXFORD | Institute for New Economic Thinking AT THE OXFORD MARTIN SCHOOL

Take-aways for the forecaster:

Analytic results: trade-off at $\psi^2 = 1$ regardless of breaks. $\therefore \alpha = 16\%$ for $N = 2$ in all settings.

Breaks in form of location shifts dominate with selection decision of second order importance. Essential to handle breaks to avoid forecast failure.

Selection is not costly – when unknown future xs similar MSFE to **known DGP**.

Simulation evidence suggest pooling works well across many settings: combination across 'non-poisonous methods' provides insurance policy. But even methods not nesting DGP also performed well.

For practitioners uncertain of the nature of the unknown DGP, a moderate selection significance level of $\alpha = 10\%–16\%$ insures against the extremes, although there will be cases when such a choice is not optimal, and updating will reveal these.

References I

Akaike, A. (1973).
Information theory and an extension of the maximum likelihood principle.
In B. N. Petrov and F. L. Csaki (Eds.), *Second International Symposium of Information Theory*, pp. 267–281. Budapest: Akademiai Kiado.

Bergmeir, C. and J. M. Hyndman, R. J. Benítez (2016).
Bagging exponential smoothing methods using stl decomposition and box–cox transformation.
International Journal of Forecasting 32, 303–312.

Campos, J., D. F. Hendry, and H.-M. Krolzig (2003).
Consistent model selection by an automatic *Gets* approach.
Oxford Bulletin of Economics and Statistics 65, 803–819.

Clements, M. P. and D. F. Hendry (2001).
Explaining the results of the M3 forecasting competition.
International Journal of Forecasting 17, 550–554.

Dantas, M. C. and F. L. C. Oliviera (2018).
Improving time series forecasting: An approach combining bootstrap aggregation, clusters and exponential smoothing.
International Journal of Forecasting 34, 748–761.

Hyndman, R. J. and B. Billah (2003).
Unmasking the theta method.
International Journal of Forecasting 19, 287–290.

Johnson, N. L. and S. Kotz (1970).
Distributions in Statistics.
New York: John Wiley.

References II

Leeb, H. and B. M. Pötscher (2009).
Model selection.
In T. Andersen, R. A. Davis, J.-P. Kreiss, and T. Mikosch (Eds.), *Handbook of Financial Time Series*, pp. 889–926. Berlin: Springer.

Pötscher, B. M. (1991).
Effects of model selection on inference.
Econometric Theory 7, 163–185.

Bergmeir and Hyndman (2016) – Bagging:

1. Box–Cox transformation with $\lambda \in [0, 1]$.
2. Decomposition into trend and remainder using LOESS.
3. Create M remainder series using moving-block bootstrap, add the trend back in, and undo the Box–Cox transformation.
4. Construct M forecasts using exponential smoothing (using AIC to select from all available models, called ETS).
5. Output the median forecast.

Their method improves on ETS in M3 on all frequencies (yearly, quartely, monthly).
But Hyndman and Billah (2003): Theta also an exponential smoothing method, and bagging only improves on Theta(2) in monthly data.

	Yearly		Quarterly		Monthly	
M3	sMAPE	MAPE	sMAPE	MAPE	sMAPE	MAPE
Theta(2).log	16.07	2.69	9.14	1.10	13.57	0.85
Theta(2)	16.72	2.77	9.24	1.12	13.91	0.86
Theta(2)*	16.97	2.81	8.96	1.09	13.89	0.86
Bagging*	17.89	3.15	10.13	1.22	13.64	0.85
Bagging2*	17.56	2.93	9.89	1.17	13.62	0.84

* results taken from published papers.
Dantas and Oliviera (2018) extends to involve clustering (Bagging2)
[M4 competition: ranked 19th, just above Theta at 20th. Card uniformly better.]

M4	Y	Q	M	W	D	H	Y	Q	M	W	D	H
	sMAPE						MASE					
Card	13.91	10.00	12.78	6.74	3.05	8.91	3.26	1.16	0.93	2.30	3.28	0.80
Theta(2).log	13.30	10.13	13.05	7.86	3.04	18.25	2.99	1.19	0.97	2.54	3.25	2.48
Bagging2	14.75	10.25	13.46	8.87	3.25	16.94	3.29	1.17	0.95	2.53	3.43	1.60

- Retention Rate: $\widetilde{\mathsf{p}}_i = \frac{1}{M}\sum_{j=1}^{M} 1_{(\widetilde{\beta}_{i,j}\neq 0)}, \quad i = 1, \ldots, N$.
- Gauge: $\frac{1}{N-n}\sum_{i=n+1}^{N} \widetilde{\mathsf{p}}_i$
- Potency: $\frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathsf{p}}_i$

$y_t = \beta'\mathbf{x}_t + \nu_t$ where $\nu_t \sim \mathsf{IN}\left[0, \sigma^2_\nu\right]$ for $t = 1, \ldots, T$ and $\mathbf{x}_t \sim \mathsf{IN}_k\left[0, \Omega\right]$.

Minimum 1-step MSFE when β and $\mathbf{x}_{T+1}$ known is conditional expectation: $\mathsf{E}\left[\nu^2_{T+1}|\mathbf{x}_{T+1}\right] = \sigma^2_\nu$.

When β estimated; $\widehat{\beta} \sim \mathsf{N}_k\left[\beta, \sigma^2_\nu\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]$, $\widehat{\sigma^2_\nu} \sim \sigma^2_\nu \frac{\chi^2_{T-1}}{(T-k)}$:
$\frac{T\widehat{\beta}'\widehat{\boldsymbol{\Omega}}\widehat{\beta}}{k\widehat{\sigma^2_\nu}} \sim \mathsf{F}^k_{T-k}\left(\psi^2_{\beta=\mathbf{0}}\right)$ with $\psi^2_{\beta=\mathbf{0}} = \frac{T\beta'\boldsymbol{\Omega}\beta}{\sigma^2_\nu}$.

Replace $\widehat{\boldsymbol{\Omega}}$ with $\boldsymbol{\Omega} = \mathsf{E}\left[\widehat{\boldsymbol{\Omega}}\right]$ and for $T > k + 2$, see Johnson and Kotz (1970, Ch.30): $\mathsf{E}\left[\mathsf{F}^k_{T-k}\left(\psi^2_{\beta=\mathbf{0}}\right)\right] = \frac{(T-k)\left(k+\psi^2_{\beta=\mathbf{0}}\right)}{k(T-k-2)} \simeq 1 + \frac{\psi^2_{\beta=\mathbf{0}}}{k}$

When $k = 1$, $\frac{T\widehat{\beta}^2\widehat{\sigma}^2_x}{\widehat{\sigma}^2_\nu} = \mathsf{t}^2_{T-1}\left(\cdot\right)$, so:

$$\mathsf{E}\left[\mathsf{t}^2_{T-1}\left(\psi^2_{\beta=\mathbf{0}}\right)\right] > c^2 \implies 1 + \psi^2_{\beta=\mathbf{0}} > c^2.$$

Let $\beta' = (\beta_1' : \beta_2')$ and $\mathbf{x}_t' = \left(\mathbf{x}_{1,t}' : \mathbf{x}_{2,t}'\right)$.
Look at relative loss between inclusion and exclusion of $\mathbf{x}_{2,t}$.
Relative loss defined by difference in conditional MSFE relative to innovation variance: $\mathsf{R}_{l(\widetilde{\nu},\widehat{\nu},1)} = \frac{\left(\mathsf{E}\left[\widetilde{\nu}_{T+1}^2|\mathcal{I}_T\right] - \mathsf{E}\left[\widehat{\nu}_{T+1}^2|\mathcal{I}_T\right]\right)}{\sigma_\nu^2}$

$\mathsf{F}_{T-k}^{k_2}$-test of $\beta_2 = \mathbf{0}$ has non-centrality parameter $\psi_{\beta_2=\mathbf{0}}^2 = \frac{T\beta_2'\mathbf{\Omega}_{22.1}\beta_2}{\sigma_\nu^2}$
such that: $\mathsf{R}_{l(\widetilde{\nu},\widehat{\nu},1)} = T^{-1}k_2\left(\left(1 + T^{-1}k_1\right)\Psi_{\beta_2=0}^2 - 1\right)$.

When $k_2 = 1$: $\mathsf{R}_{l(\widetilde{\nu},\widehat{\nu},1)} \simeq T^{-1}\left(\psi_{\beta_2=\mathbf{0}}^2 - 1\right)$.

If non-centrality $\psi_{\beta_2=\mathbf{0}}^2 > 1$ or expected $\mathrm{t}^2 > 2$ improved forecast accuracy from inclusion.
Back

OXFORD MARTIN SCHOOL UNIVERSITY OF OXFORD Institute for New Economic Thinking AT THE OXFORD MARTIN SCHOOL

DGP:

$$y_t = \beta_0 + \beta_y y_{t-1} + \beta' \mathbf{x}_t + \epsilon_t, \qquad \epsilon_t \sim \mathsf{IN}\left(0, \sigma_\epsilon^2\right)$$

$$\underset{(N\times 1)}{\mathbf{x}_t} = \begin{cases} \iota + \lambda \mathbf{x}_{t-1} + \eta_t & \text{for } t = 1, \ldots, T \\ (\iota + \nu \nabla \iota) + (\lambda + \nu \nabla \lambda)\, \mathbf{x}_{t-1} + \eta_t & \text{for } t = T+1, T+2 \end{cases}$$

$$\eta_t \sim \mathsf{IN}_N\left[\mathbf{0}, \mathbf{I}\right]$$

$\sigma_\epsilon^2 = 1$, $\beta_0 = 5$ $\beta_y = 0.5$, $N = 15$ and $n =$ no. relevant variables

- ν for shift in relevant, irrelevant, or all regressors.
- $\iota = \mathbf{1}_N$ $\nabla\iota$: 4σ mean shift in $\mathbf{x}_t$ at $T+1$.
- $\lambda = 0.5\mathbf{I}_N$ and $\nabla\lambda = 0.45$: persistence increases from 0.5 to 0.95.

OXFORD MARTIN SCHOOL | UNIVERSITY OF OXFORD | Institute for New Economic Thinking AT THE OXFORD MARTIN SCHOOL

DGP:

$$y_t = \beta_0 + \beta_y y_{t-1} + \beta' \mathbf{x}_t + \epsilon_t, \qquad \epsilon_t \sim \mathsf{IN}\left(0, \sigma_\epsilon^2\right)$$

$$\underset{(N\times 1)}{\mathbf{x}_t} = \begin{cases} \iota + \lambda \mathbf{x}_{t-1} + \eta_t & \text{for } t = 1, \ldots, T \\ (\iota + \nu \nabla \iota) + (\lambda + \nu \nabla \lambda)\,\mathbf{x}_{t-1} + \eta_t & \text{for } t = T+1, T+2 \end{cases}$$

$$\eta_t \sim \mathsf{IN}_N\left[\mathbf{0}, \mathbf{I}\right]$$

$\sigma_\epsilon^2 = 1$, $\beta_0 = 5$ $\beta_y = 0.5$, $N = 15$ and $n =$ no. relevant variables

- ν for shift in relevant, irrelevant, or all regressors.
- $\iota = \mathbf{1}_N$ $\nabla\iota$: 4σ mean shift in $\mathbf{x}_t$ at $T+1$.
- $\lambda = 0.5\mathbf{I}_N$ and $\nabla\lambda = 0.45$: persistence increases from 0.5 to 0.95.

Three experiments, $N = 15$, $n = 5$ or 8:

$$\underset{(N\times 1)}{\psi} = \begin{cases} (0,0,0,0,0,0,0,0.5,1,1.5,2,2.5,3,3.5,4)'. \\ (0,0,0,0,0,0,0,0,0,0,4,4,4,4,4)'. \\ (0,0,0,0,0,0,0,0,0,0,1,1,1,1,1)'. \end{cases}$$

GUM:

$$y_t = \overline{\beta}_0 + \beta_y y_{t-1} + \sum_{i=0}^{1} \sum_{j=1}^{N} \beta_{ij} x_{j,t-i} + \epsilon_t$$

Selection by *Autometrics* for $\alpha = (0.001, 0.01, 0.05, 0.1, 0.16, 0.32, 0.5)$
$T = 100$, $M = 1,000$, 1-step MSFEs for $y_{T+1|T}$ and $y_{T+2|T+1}$.

GUM:

$$y_t = \overline{\beta}_0 + \beta_y y_{t-1} + \sum_{i=0}^{1} \sum_{j=1}^{N} \beta_{ij} x_{j,t-i} + \epsilon_t$$

Selection by *Autometrics* for $\alpha = (0.001, 0.01, 0.05, 0.1, 0.16, 0.32, 0.5)$
$T = 100$, $M = 1,000$, 1-step MSFEs for $y_{T+1|T}$ and $y_{T+2|T+1}$.

A range of forecasting models used:

- known future exogenous regressors as infeasible benchmark
- unknown future exogenous regressors, forecasts obtained from:
  - in-sample mean;
  - selected model from ADL GUM for exogenous regressors;
  - robust forecasting devices including RW and RW with difference;
  - AR(1);
  - univariate forecasts of y_{T+h} using RW or AR(1); and
  - pooling various forecasting models.

GUM:

$$y_t = \overline{\beta}_0 + \beta_y y_{t-1} + \sum_{i=0}^{1} \sum_{j=1}^{N} \beta_{ij} x_{j,t-i} + \epsilon_t$$

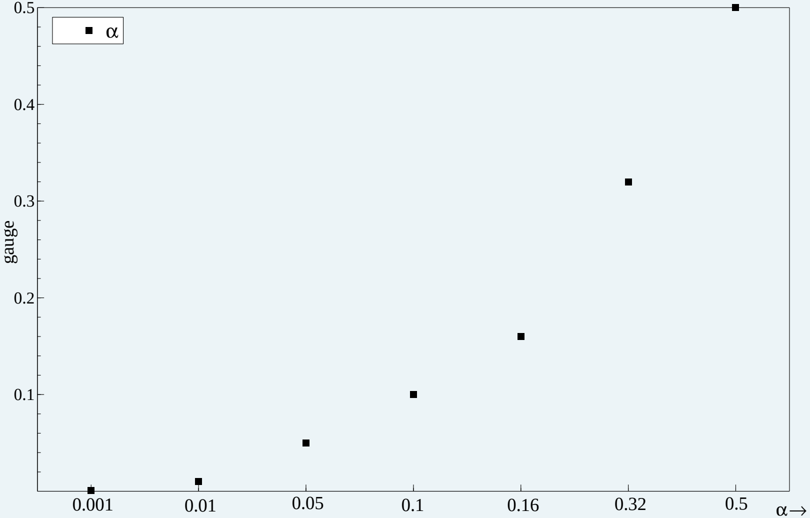
Selection by *Autometrics* for $\alpha = (0.001, 0.01, 0.05, 0.1, 0.16, 0.32, 0.5)$
$T = 100$, $M = 1,000$, 1-step MSFEs for $y_{T+1|T}$ and $y_{T+2|T+1}$.

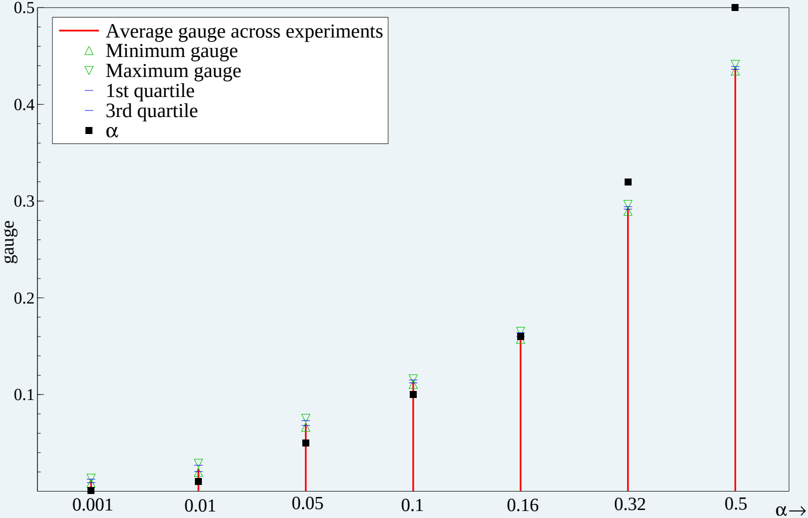
A range of forecasting models used:

- known future exogenous regressors as infeasible benchmark
- unknown future exogenous regressors, forecasts obtained from:
    - in-sample mean;
    - selected model from ADL GUM for exogenous regressors;
    - robust forecasting devices including RW and RW with difference;
    - AR(1);
    - univariate forecasts of y_{T+h} using RW or AR(1); and
    - pooling various forecasting models.

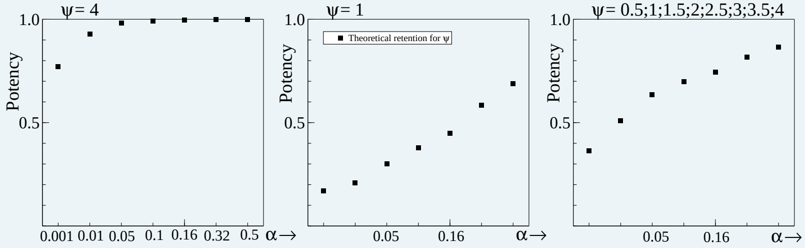
Results:

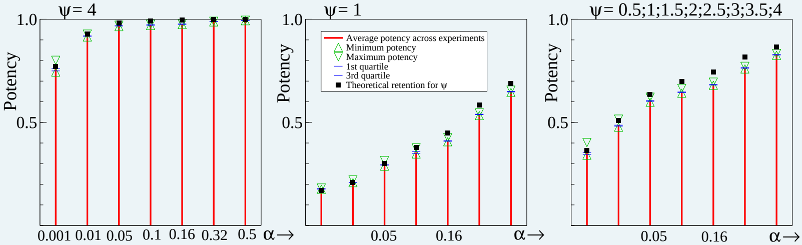
2142 distinct MSFE observations, $\overline{\text{MSFE}} = 5.15$ and $\sigma = 7.50$.

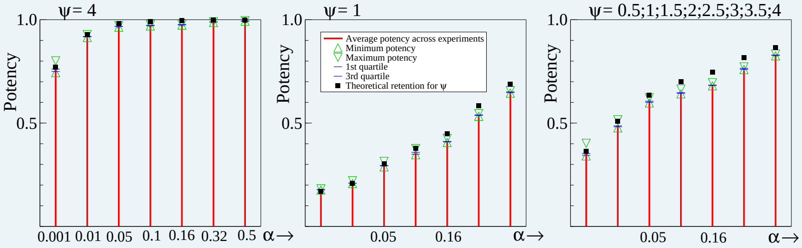

α
gauge
0.5
0.4
0.3
0.2
0.1
0.001
0.01
0.05
0.1
0.16
0.32
0.5
α→



Average gauge across experiments
Minimum gauge
Maximum gauge
1st quartile
3rd quartile
α
gauge
0.5
0.4
0.3
0.2
0.1
0.001
0.01
0.05
0.1
0.16
0.32
0.5
α→


ψ= 4
1.0
0.5
Potency
0.001 0.01 0.05 0.1 0.16 0.32 0.5 α→
ψ= 1
Theoretical retention for ψ
0.05 0.16 α→
ψ= 0.5;1;1.5;2;2.5;3;3.5;4
0.05 0.16 α→


ψ= 4
ψ= 1
ψ= 0.5;1;1.5;2;2.5;3;3.5;4
1.0
0.5
Potency
0.001 0.01 0.05 0.1 0.16 0.32 0.5 α→
0.05 0.16 α→
Average potency across experiments
Minimum potency
Maximum potency
1st quartile
3rd quartile
Theoretical retention for ψ



Null rejection frequency close to α and non-null rejections close to powers of one-off t-tests with same ψ.
$\therefore$ use *Autometrics* to evaluate theoretical results by simulation, without concern that selection algorithm influences results relative to single t-test approach.

All experiments [3 ψ's, breaks out/end-of-sample, no breaks/breaks in relevant/irrelevant/all regressors]

Rankings at $\alpha = 10\%$: (1 = smallest MSFE, 8 = largest MSFE ranking)

1. Forecast pooling over: selection, RW for $\mathbf{x}$, direct AR(1) for y
2. Direct AR(1) forecast for y
3. RW robust forecast for $\mathbf{x}$
4. Selecting from ADL GUM for $\mathbf{x}$
5. In-sample mean for $\mathbf{x}$
6. RW with difference robust forecast for $\mathbf{x}$
7. Direct RW forecast for y
8. AR(1) forecast for $\mathbf{x}$

All experiments [3 ψ's, breaks out/end-of-sample, no breaks/breaks in relevant/irrelevant/all regressors]

Rankings at $\alpha = 10\%$: (1 = smallest MSFE, 8 = largest MSFE ranking)

1. Forecast pooling over: selection, RW for $\mathbf{x}$, direct AR(1) for y
2. Direct AR(1) forecast for y
3. RW robust forecast for $\mathbf{x}$
4. Selecting from ADL GUM for $\mathbf{x}$
5. In-sample mean for $\mathbf{x}$
6. RW with difference robust forecast for $\mathbf{x}$
7. Direct RW forecast for y
8. AR(1) forecast for $\mathbf{x}$

- In-sample mean for $\mathbf{x}$: Worst model for end-of-sample breaks in relevant/all regressors but best out-of-sample.
- RW with difference robust forecast for $\mathbf{x}$: Best for end-of-sample breaks in relevant/all regressors, poor if breaks out-of-sample.
- Direct AR(1) forecast for y: best if breaks in irrelevant variables.

Features that matter across specifications:

- **potency**
- **gauge**
- **theoretical retention probability given ψ**

- Small variation in MSFE across α relative to variation across break types/DGP designs.
- Too tight or too loose α (0.1% or 50%) can worsen MSFE substantially.
- Selection at 5% preferred for $\psi = 4$, but 16% often dominates for $\psi = 1$ or mixed ψ.
- Choice of α interacts with whether break occurs in the relevant or irrelevant regressors.
- Knowing the DGP only preferred in 4 of 14 cases, irrespective of ψ, although also knowing future values of regressors (and hence breaks) always dominates.

Table below summarises MSFEs for $\psi^2_{\beta_2} = 0$ and $\psi^2_{\beta_2} = 16$
$[\sigma^2_{11} = \sigma^2_{22} = \sigma^2_{\epsilon} = 1,\ \beta_0 = 5,\ \beta_1 = 1,\ \mu_1 = \mu_2 = 2,\ \nabla\mu_2 = 4,\ \rho = 0.5]$

- Costs of selection are usually small, irrespective of ψ_{β_2}.
- Model selection reduces risk relative to worst model.
- Costs of unmodelled shifts are large, up to almost 8-fold greater than baseline stationary case.
- Even facing breaks, trade-off for selecting variables in forecasting models (retain if $\psi > 1$) still applies $\Longrightarrow$ looser significance levels than typically used.
- But when many $\beta_{2,i} = 0$ subject to location shifts, erroneously including $\mathbf{x}_2$ in model costly. Loose significance levels increase the chance that irrelevant variables with $\psi_{\beta_{2,i}} = 0$ are retained by chance significance for a given draw.

Break type & case		Out: $T+1\|T$			In: $T+2\|T+1$		
		DGP	$\alpha=0.05$	$\alpha=0.16$	DGP	$\alpha=0.05$	$\alpha=0.16$
No break							
	(i) known	1.13	1.38	1.40	1.09	1.30	1.36
	(ii) sample mean	1.59	1.63	1.64	1.57	1.59	1.62
	(iv,b) RW with diff.	2.12	2.26	2.30	2.02	2.11	2.18
	(vii) pooling		1.52	1.52		1.49	1.53
Break Relevant							
ι	(i) known	1.55	2.55	2.22	1.66	2.79	2.45
	(ii) sample mean	17.56	17.58	17.54	39.50	40.40	41.28
	(iv,b) RW with diff.	18.61	18.77	18.96	**2.53**	3.91	3.46
	(vii) pooling		17.75	17.79		17.99	16.54
λ	(i) known	1.25	1.67	1.59	1.33	1.92	1.78
	(ii) sample mean	4.46	4.51	4.50	11.96	12.12	12.35
	(iv,b) RW with diff.	4.38	4.62	4.66	**2.43**	3.40	3.11
	(vii) pooling		4.16	4.16		7.07	6.76
Break Irrelevant							
ι	(i) known	1.13	1.55	1.70	1.09	1.61	1.84
	(ii) sample mean	1.59	1.63	1.64	1.57	1.59	1.60
	(iv,b) RW with diff.	2.11	2.25	2.31	2.01	2.21	2.30
	(vii) pooling		1.52	1.54		1.54	1.57
λ	(i) known	1.13	1.44	1.49	1.09	1.39	1.56
	(ii) sample mean	1.59	1.63	1.64	1.57	1.59	1.61
	(iv,b) RW with diff.	2.12	2.25	2.31	2.02	2.11	2.20
	(vii) pooling		1.52	1.53		1.51	1.55
Break All							
ι	(i) known	1.54	2.73	2.50	1.66	2.84	2.67
	(ii) sample mean	17.88	17.90	17.86	40.01	40.93	41.69
	(iv,b) RW with diff.	18.86	18.99	19.12	**2.53**	3.76	3.46
	(vii) pooling		18.02	18.00		17.30	15.50
λ	(i) known	1.25	1.71	1.67	1.33	1.92	1.89
	(ii) sample mean	4.50	4.55	4.55	12.06	12.23	12.45
	(iv,b) RW with diff.	4.40	4.63	4.68	**2.42**	3.37	3.15
	(vii) pooling		4.19	4.18		6.99	6.64

Table: Simulation summary for 8 relevant variables with non-centralities of $0.5; 1; 1.5; 2; 2.5; 3; 3.5; 4$ and 7 irrelevant variables. Shaded cells indicate minimum MSFE for selection across methods listed; bold where knowing the DGP, but not the future values of the regressors, would have dominated.

Break type & case		Out: $T+1\|T$			In: $T+2\|T+1$		
		DGP	$\alpha=0.05$	$\alpha=0.16$	DGP	$\alpha=0.05$	$\alpha=0.16$
No break							
	(i) known	1.09	1.24	1.33	1.06	1.19	1.30
	(ii) sample mean	1.90	1.96	1.97	1.89	1.93	1.95
	(iv,b) RW with diff.	2.59	2.68	2.77	2.51	2.62	2.71
	(vii) pooling		1.73	1.77		1.76	1.80
Break Relevant							
ι	(i) known	1.32	1.62	1.69	1.43	1.78	1.86
	(ii) sample mean	21.22	21.20	21.22	47.18	48.42	49.52
	(iv,b) RW with diff.	22.60	22.77	22.78	**2.92**	3.37	3.41
	(vii) pooling		21.51	21.56		20.13	19.51
λ	(i) known	1.16	1.34	1.40	1.19	1.43	1.51
	(ii) sample mean	5.46	5.52	5.54	14.52	14.80	15.11
	(iv,b) RW with diff.	5.17	5.35	5.44	**2.87**	3.20	3.28
	(vii) pooling		4.91	4.96		8.04	7.99
Break Irrelevant							
ι	(i) known	1.09	1.45	1.70	1.06	1.67	2.01
	(ii) sample mean	1.89	1.95	1.96	1.88	1.92	1.94
	(iv,b) RW with diff.	2.57	2.67	2.75	2.49	2.73	2.90
	(vii) pooling		1.74	1.78		1.82	1.94
λ	(i) known	1.09	1.30	1.42	1.06	1.35	1.48
	(ii) sample mean	1.90	1.96	1.97	1.89	1.93	1.95
	(iv,b) RW with diff.	2.58	2.67	2.77	2.51	2.62	2.65
	(vii) pooling		1.73	1.77		1.79	1.84
Break All							
ι	(i) known	1.32	1.79	2.03	1.43	2.04	2.41
	(ii) sample mean	21.88	21.88	21.89	48.27	49.58	50.64
	(iv,b) RW with diff.	23.14	23.30	23.32	**2.94**	3.43	3.59
	(vii) pooling		22.10	22.14		18.92	17.70
λ	(i) known	1.16	1.39	1.49	1.19	1.52	1.66
	(ii) sample mean	5.55	5.60	5.64	14.71	15.01	15.28
	(iv,b) RW with diff.	5.21	5.40	5.47	**2.87**	3.17	3.29
	(vii) pooling		4.98	5.01		7.88	7.77

Table: Simulation summary for 5 relevant variables with non-centralities of 4 and 10 irrelevant variables. Shaded cells indicate minimum MSFE for selection across methods listed.

Break type & case		Out: $T+1\|T$			In: $T+2\|T+1$		
		DGP	$\alpha=0.05$	$\alpha=0.16$	DGP	$\alpha=0.05$	$\alpha=0.16$
No break							
	(i) known	1.10	1.24	1.31	1.07	1.22	1.30
	(ii) sample mean	1.06	1.09	1.09	1.06	1.08	1.08
	(iv,b) RW with diff.	1.19	1.29	1.37	1.20	1.23	1.33
	(vii) pooling		1.11	1.12		1.08	1.10
Break Relevant							
ι	(i) known	1.32	2.02	1.89	1.44	2.49	2.07
	(ii) sample mean	2.32	2.34	2.34	4.07	4.22	4.26
	(iv,b) RW with diff.	2.52	2.61	2.73	**1.53**	2.60	2.18
	(vii) pooling		2.39	2.42		3.09	2.73
λ	(i) known	1.16	1.42	1.45	1.21	1.64	1.62
	(ii) sample mean	1.30	1.33	1.33	1.88	1.93	1.94
	(iv,b) RW with diff.	1.38	1.53	1.60	**1.28**	1.70	1.63
	(vii) pooling		1.35	1.36		1.70	1.62
Break Irrelevant							
ι	(i) known	1.10	1.41	1.66	1.07	1.60	1.93
	(ii) sample mean	1.06	1.09	1.09	1.06	1.07	1.08
	(iv,b) RW with diff.	1.19	1.28	1.37	1.20	1.40	1.62
	(vii) pooling		1.11	1.12		1.10	1.14
λ	(i) known	1.10	1.28	1.38	1.07	1.37	1.49
	(ii) sample mean	1.06	1.09	1.09	1.06	1.08	1.08
	(iv,b) RW with diff.	1.19	1.29	1.38	1.20	1.27	1.34
	(vii) pooling		1.11	1.12		1.10	1.10
Break All							
ι	(i) known	1.32	2.23	2.15	1.44	2.71	2.59
	(ii) sample mean	2.36	2.38	2.38	4.14	4.28	4.34
	(iv,b) RW with diff.	2.56	2.66	2.74	**1.53**	2.71	2.33
	(vii) pooling		2.43	2.45		3.05	2.67
λ	(i) known	1.16	1.47	1.49	1.21	1.74	1.74
	(ii) sample mean	1.31	1.34	1.34	1.90	1.94	1.95
	(iv,b) RW with diff.	1.39	1.53	1.59	**1.28**	1.72	1.66
	(vii) pooling		1.35	1.36		1.70	1.61

Table: Simulation summary for 5 relevant variables with non-centralities of 1 and 10 irrelevant variables. Shaded cells indicate minimum MSFE for selection across methods listed.

			(ii)	(iii)	(iva)	(ivb)	(v)	(via)	(vib)	(vii)
No Break										
	(1)	Out	3	4	5	7	*8*	6	**1**	2
		In	3	4	5	7	*8*	6	**1**	2
	(2)	Out	4	3	5	7	*8*	6	2	**1**
		In	3	4	5	7	*8*	6	2	**1**
	(3)	Out	2	4	5	6	*8*	7	**1**	3
		In	2	4	5	6	*8*	7	**1**	3
Break Relevant										
	(1)	Out	1	3	6	7	*8*	5	2	4
ι		In	*8*	3	2	**1**	6	5	7	4
	(2)	Out	1	4	5	7	*8*	6	2	3
		In	*8*	4	2	**1**	6	5	7	3
	(3)	Out	1	4	5	6	*8*	7	2	3
		In	*8*	4	2	**1**	7	3	6	5
	(1)	Out	5	2	4	7	*8*	6	3	**1**
λ		In	*8*	4	2	**1**	6	5	7	3
	(2)	Out	7	3	2	6	*8*	5	4	**1**
		In	*8*	4	2	**1**	6	5	7	3
	(3)	Out	2	4	5	6	*8*	7	**1**	3
		In	7	4	3	2	*8*	5	6	**1**
Break Irrelevant										
	(1)	Out	3	4	5	7	*8*	6	**1**	2
ι		In	3	6	4	7	*8*	5	**1**	2
	(2)	Out	3	4	5	7	*8*	6	2	**1**
		In	3	6	5	7	*8*	4	**1**	2
	(3)	Out	2	5	4	6	*8*	7	**1**	3
		In	2	6	4	7	*8*	5	**1**	3
	(1)	Out	3	4	5	7	*8*	6	2	**1**
λ		In	3	4	5	7	*8*	6	**1**	2
	(2)	Out	4	3	5	7	*8*	6	2	**1**
		In	3	6	4	7	*8*	5	**1**	2
	(3)	Out	2	4	5	6	*8*	7	**1**	3
		In	2	5	4	6	*8*	7	**1**	3
Break All										
	(1)	Out	1	4	5	7	*8*	6	2	3
ι		In	*8*	3	2	**1**	6	5	7	4
	(2)	Out	1	4	5	7	*8*	6	2	3
		In	*8*	3	2	**1**	6	5	7	4
	(3)	Out	1	4	5	6	*8*	7	2	3
		In	*8*	5	2	**1**	7	3	6	4
	(1)	Out	5	2	4	7	*8*	6	3	**1**
λ		In	*8*	3	2	**1**	6	5	7	4
	(2)	Out	7	3	2	6	*8*	5	4	**1**
		In	*8*	4	2	**1**	6	5	7	3
	(3)	Out	2	4	5	6	*8*	7	**1**	3
		In	7	4	3	2	*8*	5	6	**1**
Average			4.2	4.0	3.9	5.1	7.6	5.6	3.1	2.5

Table: Simulation summary rankings for $\alpha = 10\%$. 'Out' refers to forecasts for $T+1|T$, i.e. the break is out-of-sample. 'In' refers to forecasts for $T+2|T+1$ where the break is in-sample. (1) is for the case with $\psi = (0,0,0,0,0,0,0,0.5,1,1.5,2,2.5,3,3.5,4)'$, (2) is case $\psi = (0,0,0,0,0,0,0,0,0,0,4,4,4,4,4)'$, and (3) is for $\psi = (0,0,0,0,0,0,0,0,0,0,1,1,1,1,1)'$. Lower case Roman numerals respectively denote forecasting the unknown future exogenous regressors by: (ii) the in-sample mean; (iii) selecting from the GUM (**??**); (iva) a random walk; (ivb) that with the added difference; (v) an AR(1); (via) a direct random walk forecast of y; (vib) a direct AR(1) forecast of y; and (vii) pooling.