In-Out of Sample Fit/Qrinkage

Two Issues:

- Understand why models that fit well in-sample tend not to do well out-of-sample
- Adjust the parameter estimates prior to out-of-sample analysis – Qrinkage

In/Out of Sample

- goodness-of-fit statistics: LR_{In} and LR_{out}

$$\begin{aligned} L_{In} &\sim Z_1' Z_1 \\ LR_{out} &\sim 2Z_1' Z_2 - Z_1' Z_1 \end{aligned}$$

- LR_{In} tends to be inflated by estimation error in finite samples (overfit bias)
- LR_{In} and LR_{out} are negatively correlated

Implications

- Good in-sample fit translates into poor out-of-sample fit

Implications

- Good in-sample fit translates into poor out-of-sample fit
- Out-of-sample analysis is less likely to produce spurious results
- Evidence for out-of-sample predictability is stronger than in-sample evidence.

What's wrong with information criteria:

- Akaike's FPE:

$$\begin{aligned} E(y_{T+1} - X_T'\hat{\beta}(k))^2 &= \sigma^2 + \sigma^2 E[(\hat{\beta}(k) - \beta(k))' X_T X_T' (\hat{\beta}(k) - \beta \\ \sqrt{T}(\hat{\beta}(k) - \beta(k)) &\sim N(0, \Gamma_k^{-1}), \quad \Gamma_k = E(X_T X_T') \end{aligned}$$

What's wrong with information criteria:

- Akaike's FPE:

$$
\begin{aligned}
E(y_{T+1} - X_T'\hat{\beta}(k))^2 &= \sigma^2 + \sigma^2 E[(\hat{\beta}(k) - \beta(k))' X_T X_T' (\hat{\beta}(k) - \beta \\
\sqrt{T}(\hat{\beta}(k) - \beta(k)) &\sim N(0, \Gamma_k^{-1}), \quad \Gamma_k = E(X_T X_T') \\
FPE &= \sigma^2(1 + k/T) \\
\log FPE &= \log \hat{\sigma}_k^2 + 2k/(T-k).
\end{aligned}
$$

What's wrong with information criteria:

- Akaike's FPE:

$$
\begin{aligned}
E(y_{T+1} - X_T'\hat{\beta}(k))^2 &= \sigma^2 + \sigma^2 E[(\hat{\beta}(k) - \beta(k))' X_T X_T' (\hat{\beta}(k) - \beta \\
\sqrt{T}(\hat{\beta}(k) - \beta(k)) &\sim N(0, \Gamma_k^{-1}), \quad \Gamma_k = E(X_T X_T') \\
FPE &= \sigma^2(1 + k/T) \\
\log FPE &= \log \hat{\sigma}_k^2 + 2k/(T - k).
\end{aligned}
$$

- comparing two models with the same k amounts to comparing in sample fit

What's wrong with information criteria:

- Akaike's FPE:

$$
\begin{aligned}
E(y_{T+1} - X_T'\hat{\beta}(k))^2 &= \sigma^2 + \sigma^2 E[(\hat{\beta}(k) - \beta(k))' X_T X_T' (\hat{\beta}(k) - \beta \\
\sqrt{T}(\hat{\beta}(k) - \beta(k)) &\sim N(0, \Gamma_k^{-1}), \quad \Gamma_k = E(X_T X_T') \\
FPE &= \sigma^2(1 + k/T) \\
\log FPE &= \log \hat{\sigma}_k^2 + 2k/(T-k).
\end{aligned}
$$

- comparing two models with the same k amounts to comparing in sample fit
- note: penalty of k is not data dependent

What's wrong with information criteria:

- Akaike's FPE:

$$
\begin{aligned}
E(y_{T+1} - X_T'\hat{\beta}(k))^2 &= \sigma^2 + \sigma^2 E[(\hat{\beta}(k) - \beta(k))' X_T X_T' (\hat{\beta}(k) - \beta
\\
\sqrt{T}(\hat{\beta}(k) - \beta(k)) &\sim N(0, \Gamma_k^{-1}), \quad \Gamma_k = E(X_T X_T') \\
FPE &= \sigma^2(1 + k/T) \\
\log FPE &= \log \hat{\sigma}_k^2 + 2k/(T-k).
\end{aligned}
$$

- comparing two models with the same k amounts to comparing in sample fit
- note: penalty of k is not data dependent
- note: does not take into account what is kmax.

How to Reduce Bias from Overfitting?

How to Reduce Bias from Overfitting?

- Ridge regression: How to set shrinkage parameter?
- Qrinkage: finds the shrinkage parameter
  - orthogonal regressors $y = x\theta + \epsilon$
  - unrestricted estimates: $\hat{\theta}_i, i = 1, \ldots N$
  - find $\tilde{\theta} = \kappa\hat{\theta}$ such that $2(LR(\tilde{\theta}) - LR(\hat{\theta})) = k$

How to Reduce Bias from Overfitting?

- Ridge regression: How to set shrinkage parameter?
- Qrinkage: finds the shrinkage parameter
  - orthogonal regressors $y = x\theta + \epsilon$
  - unrestricted estimates: $\hat{\theta}_i, i = 1, \ldots N$
  - find $\tilde{\theta} = \kappa\hat{\theta}$ such that $2(LR(\tilde{\theta}) - LR(\hat{\theta})) = k$

$$\kappa_i^* = \max(0, \sqrt{\frac{\lambda_i \hat{\sigma}^2}{\delta_i^2 n}}) \approx \max(0, 1 - \frac{1}{|t_{\hat{\theta}_i}|}).$$

Relation to AIC

- $\hat{KL} = L_T(\theta_0) - L_T(\hat{\theta})$ is biased for $KL = E_0[L(\theta_0) - L(\hat{\theta})]$

Relation to AIC

- $\hat{KL} = L_T(\theta_0) - L_T(\hat{\theta})$ is biased for $KL = E_0[L(\theta_0) - L(\hat{\theta})]$
- Put $L^* = L_T + k$, then $\lim_{T\rightarrow} E_0[T(KL - KL^*)] = 0$.

Relation to AIC

- $\hat{KL} = L_T(\theta_0) - L_T(\hat{\theta})$ is biased for $KL = E_0[L(\theta_0) - L(\hat{\theta})]$
- Put $L^* = L_T + k$, then $\lim_{T\to} E_0[T(KL - KL^*)] = 0$.
- $L_T(\hat{\theta}) \propto \hat{\sigma}^2/2 \Rightarrow AIC(k) = \log(\hat{\sigma}^2) + 2k/T$.

Relation to AIC

- $\hat{KL} = L_T(\theta_0) - L_T(\hat{\theta})$ is biased for $KL = E_0[L(\theta_0) - L(\hat{\theta})]$
- Put $L^* = L_T + k$, then $\lim_{T \to} E_0[T(KL - KL^*)] = 0$.
- $L_T(\hat{\theta}) \propto \hat{\sigma}^2/2 \Rightarrow AIC(k) = \log(\hat{\sigma}^2) + 2k/T$.
- Qrinkage tries to bias correct the same objective function

- AIC performs hard thresholding (results can be unstable)
- not useful for comparing models with the same k

- AIC performs hard thresholding (results can be unstable)
- not useful for comparing models with the same k
- for orthogonal regressors, AIC $\Rightarrow$

$$\tilde{\theta}_i = \hat{\theta} \cdot I(|t_{\hat{\theta}_i}| > \sqrt{2}) = \begin{cases} \hat{\theta}_i & |t_{\hat{\theta}_i}| > \sqrt{2}| \\ 0 & \text{otherwise} \end{cases}$$

- AIC performs hard thresholding (results can be unstable)
- not useful for comparing models with the same k
- for orthogonal regressors, AIC $\Rightarrow$

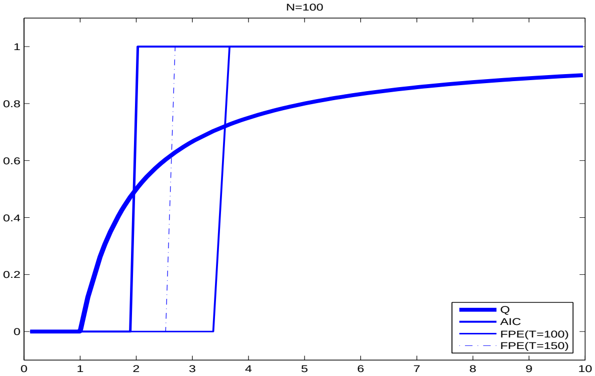
$$\tilde{\theta}_i = \hat{\theta} \cdot I(|t_{\hat{\theta}_i}| > \sqrt{2}) = \begin{cases} \hat{\theta}_i & |t_{\hat{\theta}_i}| > \sqrt{2}| \\ 0 & \text{otherwise} \end{cases}$$

- Qrinkage $\Rightarrow$

$$\theta_i = \hat{\theta}_i \cdot \max(0, 1 - 1/|t_{\hat{\theta}_i}|) = \begin{cases} \hat{\theta}_i(1 - 1/|t_{\hat{\theta}_i}|) & |t_{\hat{\theta}_i}| > 1| \\ 0 & \text{otherwise} \end{cases}$$

Qrinkage: $\tilde{\theta}_i = \hat{\theta}_i \max(0, 1 - 1/|t_i|)$

- $|t_i| = 10$, $\tilde{\theta}_i = .9\hat{\theta}_i$
- $|t_i| = 5$, $\tilde{\theta}_i = .8\hat{\theta}_i$
- $|t_i| = 2$, $\tilde{\theta}_i = .5\hat{\theta}_i$
- $|t_i| = 4/3$, $\tilde{\theta}_i = .25\hat{\theta}_i$
- $|t_i| < 1$, $\tilde{\theta}_i = 0$



N=100
1
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
6
7
8
9
10
Q
AIC
FPE(T=100)
FPE(T=150)

Applications

- Diffusion Index forecasts
  - PC + Shrinkage: two dimension reductions, Why?

Applications

- Diffusion Index forecasts
    - PC + Shrinkage: two dimension reductions, Why?
    - best principal components in x need not be best predictors for y
    - lag length of factors

Applications

- Diffusion Index forecasts
  - PC + Shrinkage: two dimension reductions, Why?
  - best principal components in x need not be best predictors for y
  - lag length of factors

- predictive regressions, with or without principal components
  - objective is to explain y, not factors that explain x.

Applications

- Diffusion Index forecasts
  - PC + Shrinkage: two dimension reductions, Why?
  - best principal components in x need not be best predictors for y
  - lag length of factors

- predictive regressions, with or without principal components
  - objective is to explain y, not factors that explain x.

- many instrument IV problems

Alternative procedures: LARS, Boosting, other data mining methods

Future Work

1. More than one way to bias correct the objective function
   - What is the true model?
   - Does the true model has a finite number of parameters?
   - Is the structure sparse?

Future Work

2. 'correct' model selection is often not the ultimate goal
   - In AR(∞) models AIC/FPE minimizes MSE
   - In AR(p) models, BIC gives consistent model selection.

Future Work

2. 'correct' model selection is often not the ultimate goal
   - In AR(∞) models AIC/FPE minimizes MSE
   - In AR(p) models, BIC gives consistent model selection.

   - consistent model selection $\not\Rightarrow$ accurate unit root test

Future Work

2. 'correct' model selection is often not the ultimate goal
   - In AR(∞) models AIC/FPE minimizes MSE
   - In AR(p) models, BIC gives consistent model selection.

   - consistent model selection $\not\Rightarrow$ accurate unit root test

   - if shrinkage parameter is tuned to give conservative model selection, estimators are uniformly $\sqrt{T}$ consistent
   - if shrinkage parameter is tuned to give consistent model selection, does not get $\sqrt{T}$ consistency

Future Work

2. 'correct' model selection is often not the ultimate goal
   - In AR(∞) models AIC/FPE minimizes MSE
   - In AR(p) models, BIC gives consistent model selection.

   - consistent model selection $\nRightarrow$ accurate unit root test

   - if shrinkage parameter is tuned to give conservative model selection, estimators are uniformly $\sqrt{T}$ consistent
   - if shrinkage parameter is tuned to give consistent model selection, does not get $\sqrt{T}$ consistency

What is the objective? One criterion fits all?

Future Work

3. Open issues for Qrinkage:
   - how to form prediction confidence interval?
   - other κ_is? kernel weighting?

Future Work

3. Open issues for Qrinkage:
   - how to form prediction confidence interval?
   - other κ_is? kernel weighting?
   - How to accommodate non- orthogonal regressors?

Future Work

3. Open issues for Qrinkage:
   - how to form prediction confidence interval?
   - other κ_is? kernel weighting?
   - How to accommodate non- orthogonal regressors?
   - still does not take into account how many models are being compared.

Nice work!