Machine Learning in Econometrics: Lecture 10

INSTRUCTOR: CHAOYI CHEN
NJE & MNB

November 7, 2023

Topic 4: Practical Weight Selection

- The theory we have described concerns the infeasible best weights
- They are unknown
- How can they be estimated? feasible approaches
  - Plug-in
  - Mallows
  - CV

- Chu-An Liu, Journal of Econometrics, 2015 (Liu 2015)
- Recall

$$
\begin{aligned}
\text{wmse}(\widehat{B}(\mathrm{w})) &= \mathrm{w}^\top \bar{M}\mathrm{w}, \\
\bar{M} &= \left[B^\top \left(X^\top A_m^\top - I\right) W \left(A_\ell X - I\right) B + \operatorname{tr}\left(A_m D A_\ell^\top W\right)\right]_{m\ell}, \\
A_m &= \begin{bmatrix} \left(X_m^\top X_m\right)^{-1} X_m^\top \\ 0 \end{bmatrix}.
\end{aligned}
$$

- Estimator

$$
\begin{aligned}
\widehat{B} &= \left(X^\top X\right)^{-1} X^\top y, \\
\widehat{M} &= \left[\widehat{B}^\top \left(X^\top A_m^\top - I\right) W \left(A_\ell X - I\right) \widehat{B} + \operatorname{tr}\left(A_m \widehat{D} A_\ell^\top W\right)\right]_{m\ell}, \\
\widehat{D} &= \operatorname{diag}\left(\widehat{e}_1^2, \ldots, \widehat{e}_n^2\right)
\end{aligned}
$$

Topic 4: Plug-in Approach

- $\widehat{M} = \left[\widehat{B}^\top \left(X^\top A_m^\top - I\right) W \left(A_\ell X - I\right) \widehat{B} + \operatorname{tr}\left(A_m \widehat{D} A_\ell^\top W\right)\right]_{m\ell}$
- $\widehat{\text{wmse}}(\widehat{B}(\mathrm{w})) = \mathrm{w}^\top \widehat{W} \mathrm{w}$
- $\widehat{\mathrm{w}} = \underset{\mathrm{w}}{\text{argmin}}\ \mathrm{w}^\top \widehat{W} \mathrm{w}$ s.t. $\sum_{m=1}^{M} \mathrm{w}_m = 1$ and $0 \leq \mathrm{w} \leq 1$
- Quadratic programming solution

Topic 4: Comments on Plug-In Approach

- Simple, computationally quick
- Works for any weight matrix W
- If W is rank one (puts rank on a single linear combination)
  - This reduces to a Focused Information Criterion
  - Hjort and Claeskens introduced this as Frequentist Model Averaging (FMA) estimator(Hjort and Claeskens 2003)
- Disadvantages
  - $\widehat{M}$ is anbiased estimator for $\bar{M}$
  - This is because when the estimated squared bias is of the same order as the variance, then the variance of the estimated bias term is of the same order
  - This bias can be corrected, but we do not pursue this here

Topic 4: Mallows Model Averaging (MMA) Criterion

- Hansen proposed the least square model averaging with Mallow criterion (Hansen 2007)
- The Mallows criterion applies to regression models with linear estimators
- Recall m^{th} regression uses a subset $\mathbf{X}_m$ of regressors

$$\mathbf{y} = \widehat{\mathbf{m}}_m + \widehat{\mathbf{e}}_m = \mathbf{X}_m \widehat{\mathbf{B}}_m + \widehat{\mathbf{e}}_m$$

$$\widehat{\mathbf{B}}_m = \left(\mathbf{X}_m^\top \mathbf{X}_m\right)^{-1} \mathbf{X}_m^\top \mathbf{y}$$

$$\widehat{\mathbf{m}}_m = \mathbf{X}_m \left(\mathbf{X}_m^\top \mathbf{X}_m\right)^{-1} \mathbf{X}_m^\top \mathbf{y} = \mathbf{P}_m \mathbf{y}$$

- Therefore we have

$$\widehat{\mathbf{m}}(\mathrm{w}) = \sum_{m=1}^{M} \mathrm{w}_m \widehat{\mathbf{m}}_m = \sum_{m=1}^{M} \mathrm{w}_m \mathbf{P}_m \mathbf{y} = \mathbf{P}(\mathrm{w}) \mathbf{y}$$

$\mathbf{P}(\mathrm{w}) = \sum_{m=1}^{M} \mathrm{w}_m \mathbf{P}_m$ is a weighted average of projection matrice

Topic 4: Mallows Criterion

- $C = \widehat{\mathbf{e}}^\top \widehat{\mathbf{e}} + 2\tilde{\sigma}^2 \text{tr}(\mathscr{A})$
- For least square averaging estimator
  - $\mathscr{A} = \mathbf{P}(\mathrm{w})$
  - $\text{tr}(\mathscr{A}) = \text{tr}(\mathbf{P}(\mathrm{w})) = \sum_{m=1}^{M} \mathrm{w}_m \text{tr}(\mathbf{P}_m) = \sum_{m=1}^{M} \mathrm{w}_m K_m$
  - K_m is number of estimated coefficients in model m.
- $C(\mathrm{w}) = \widehat{\mathbf{e}}(\mathrm{w})^\top \widehat{\mathbf{e}}(\mathrm{w}) + 2\tilde{\sigma}^2 \sum_{m=1}^{M} \mathrm{w}_m K_m$
- The penalty is the weighted average of the number of coefficients

- Stack the residual vectors by column

$$\widehat{\mathbf{E}} = [\widehat{\mathbf{e}}_1, \ldots, \widehat{\mathbf{e}}_M]$$
$$\widehat{\mathbf{e}}(\mathrm{w}) = \widehat{\mathbf{E}}\mathrm{w}$$
$$\widehat{\mathbf{e}}(\mathrm{w})^\top \widehat{\mathbf{e}}(\mathrm{w}) = \mathrm{w}^\top \widehat{\mathbf{E}}^\top \widehat{\mathbf{E}}\mathrm{w}$$

- Stack the K_m

$$\mathbf{K} = (K_1, \ldots, K_M)^\top$$
$$\sum_{m=1}^{M} \mathrm{w}_m K_m = \mathrm{w}^\top \mathbf{K}$$

- $C(\mathrm{w}) = \widehat{\mathbf{e}}(\mathrm{w})^\top \widehat{\mathbf{e}}(\mathrm{w}) + 2\tilde{\sigma}^2 \sum_{m=1}^{M} \mathrm{w}_m K_m = \boxed{\mathrm{w}^\top \widehat{\mathbf{E}}^\top \widehat{\mathbf{E}}\mathrm{w} + 2\tilde{\sigma}^2 \mathrm{w}^\top \mathbf{K}}$

Topic 4: Mallow selection

- $C(\mathrm{w}) = \mathrm{w}^\top \widehat{\mathbf{E}}^\top \widehat{\mathbf{E}} \mathrm{w} + 2\tilde{\sigma}^2 \mathrm{w}^\top \mathbf{K}$ is a quadratic c function in weight vector w

- Mallows selection
  - $\widehat{\mathrm{w}} = \underset{\mathrm{w}}{\mathrm{argmin}}\ C(\mathrm{w})$ s.t. $\sum_{m=1}^{M} \mathrm{w}_m = 1$ and $0 \leq \mathrm{w} \leq 1$
  - Quadratic Programming solution
  - You can use `quadprog` in R

- Given selected weights $\widehat{\mathrm{w}}$
  - $\widehat{\mathbf{B}} = \widehat{\mathbf{B}}(\widehat{\mathrm{w}}) = \sum_{m=1}^{M} \widehat{\mathrm{w}}_m \begin{bmatrix} \widehat{\mathbf{B}_m} \\ 0 \end{bmatrix}$
  - Weighted average of least squares estimates using selected weights
  - Generalization of model selection, where $\widehat{\mathrm{w}}_m = \{0, 1\}$

- The quadratic programming solution is quick and reliable even with hundreds of models

- Mallows criterion is unbiased for regression fit (Mallows Theorem)

Topic 4: Computation in R

- You need the quadprog package installed on the computer
- `library(quadprog)`
- `quadprog` solves the minimizes functions of the form $-d^\top b + (1/2) b^\top D b$ s.t. $A^\top b \geq b_0$ and equality constraint
- In our notation
  - $b = \mathrm{w}$
  - $d = -\tilde{\sigma}^2 \mathbf{K}$
  - $D = \widehat{\mathbf{E}}^\top \widehat{\mathbf{E}}$
  - $M = \#$ of models
- Command
  - `QP <- solve.QP(Dmat,dvec,Amat,bvec,1)`
  - `Dmat <- t(e)%*%e` where e=$n \times M$ M matrix of residuals from M models
  - `dvec <- K*sig2` $M \times 1$ 1 vector of number of regression parameters in each model
  - The "1" says that the first constraint is an equality, the remainder inequality

Topic 4: Impose Contraints

- The technical issue is to construct Amat and bvec to impose constraints of the form $A^\top b \geq b_0$
- The first constraint is that the sum of the weights to 1
- The second set of constraints is that the weights are greater than zero
- The third set of constraints is that the weights are less than one
- `Amat<-t(rbind(matrix(1,nrow=1,ncol=M),diag(M),-diag(M)))`
- `bvec<-rbind(1,matrix(0,nrow=M,ncol=1),matrix(-1,nrow=M,ncol=1))`
- `QP <- solve.QP(Dmat,dvec,Amat,bvec,1)`
- `w <- QP$solution`

Topic 4: Jackknife Model Averaging (JMA) Criterion

- Hansen and Racine 2012 Journal of Econometrics (Hansen and Racine 2012)
- The Mallows criterion applies to regression models
- Now, we also allow for the Heteroskedasticity
- Averaging estimator of conditional mean at i^{th} observation

$$\widehat{m}_i(\mathrm{w}) = \sum_{m=1}^{M} \mathrm{w}_m \mathbf{x}_{mi}^{\top} \widehat{\mathbf{B}}_m$$

- The leave one out estimator is

$$\tilde{m}_i(\mathrm{w}) = \sum_{m=1}^{M} \mathrm{w}_m \mathbf{x}_{mi}^{\top} \widehat{\mathbf{B}}_{(-i)m}$$

- Prediction error

$$\tilde{e}_i(\mathrm{w}) = y_i - \tilde{m}_i(\mathrm{w})$$

Topic 4: CV/JMA criterion

- $\tilde{e}_i(\mathrm{w}) = y_i - \tilde{m}_i(\mathrm{w}) = y_i - \sum_{m=1}^{M} \mathrm{w}_m \mathbf{x}_{mi}^{\top} \widehat{\mathbf{B}}_{(-i)m}$
- $\mathrm{CV}(\mathrm{w}) = \tilde{\mathbf{e}}(\mathrm{w})^{\top} \tilde{\mathbf{e}}(\mathrm{w})$
- JMA select
  - $\widehat{\mathrm{w}} = \underset{\mathrm{w}}{\operatorname{argmin}}\ \mathrm{CV}(\mathrm{w})$ s.t. $\sum_{m=1}^{M} \mathrm{w}_m = 1$ and $0 \leq \mathrm{w} \leq 1$
  - Properties are similar to Mallows select
  - Asymptotic optimality holds under conditional heteroskedasticity

- Stack the prediction error vectors by column
  - $\tilde{\mathbf{e}}_m = (\tilde{e}_{1,m}, \ldots, \tilde{e}_{1,M})$
  - $\tilde{\mathbf{E}} = [\tilde{\mathbf{e}}_1, \ldots, \tilde{\mathbf{e}}_M]$
  - $\tilde{\mathbf{e}}(\mathrm{w}) = \tilde{\mathbf{E}}\mathrm{w}$
  - $\tilde{\mathbf{e}}(\mathrm{w})^\top \tilde{\mathbf{e}}(\mathrm{w}) = \mathrm{w}^\top \tilde{\mathbf{E}}^\top \tilde{\mathbf{E}}\mathrm{w}$
- $\mathrm{CV}(\mathrm{W}) = \mathrm{w}^\top \tilde{\mathbf{E}}^\top \tilde{\mathbf{E}}\mathrm{w}$
  - Quadratic function in weight vector w
  - Minimization is a Quadratic Programming solution
- Numerically as simple as Mallows criterion

Topic 4: quadprog

- In our notation
  - $b = \mathrm{w}$
  - $d = 0$
  - $D = \widehat{\mathbf{E}}^\top \widehat{\mathbf{E}}$
  - $M = \#$ of models
- Command
  - `QP <- solve.QP(Dmat,matrix(0,M,1),Amat,bvec,1)`
  - `Dmat <- t(pe)%*%pe` where pe$=n \times M$ M matrix of prediction errors from M models
  - Same constraints as for Mallows criterion inequality

Topic 4: Simulation Evidence

- MMA and JMA perform better than selection methods
- MMA and JMA perform better than AIC and BIC
- JMA performs better than MMA, especially under heteroskedasticity
- Improvement is not uniform in parameter space

Reference

Hansen, Bruce E. (2007). Least squares model averaging. *Econometrica* **75**(4), 1175–1189.

Hansen, Bruce E. and Jeffrey S. Racine (2012). Jackknife model averaging. *Journal of Econometrics* **167**(1), 38–46.

Hjort, Nils Lid and Gerda Claeskens (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98**(464), 879–899.

Liu, Chu-An (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics* **186**(1), 142–159.