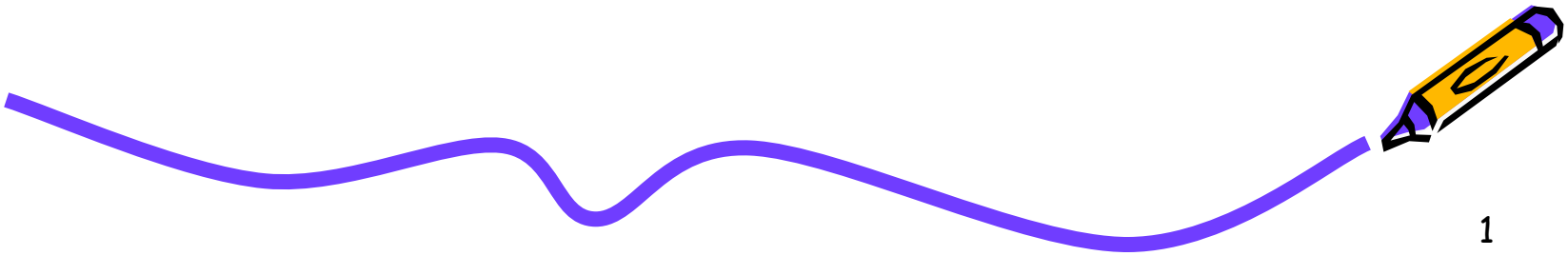
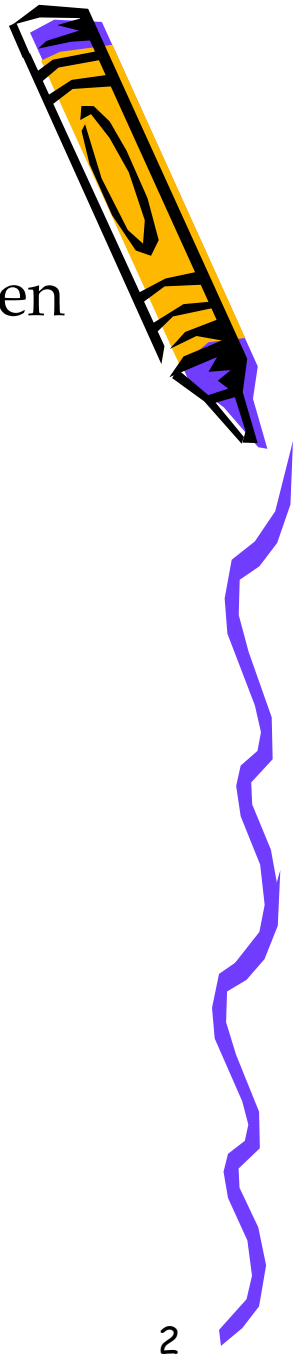
Weight Selection for a Model Average Estimator

Alan Wan

City University of Hong Kong

(joint work with H. Liang and G. Zou, University of Rochester)



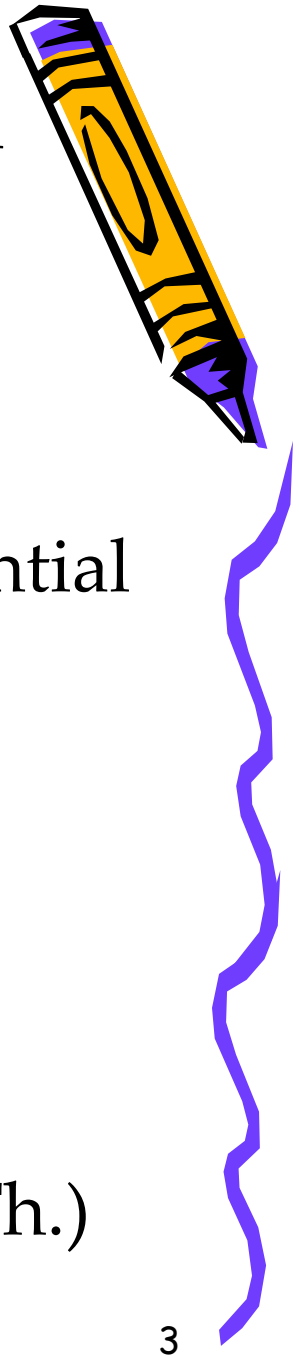


- Model selection methods assume final model chosen in advance

- Under-reporting of variability and confidence intervals.

- Papers on under-reporting due to model selection:
  Danilov and Magnus (2004, J. of Econometrics)
  Leeb and Pötscher (2006, Annals of Stat)
  Leeb and Pötscher (2008, Econometric Theory)

- Current paper – frequentist model averaging
- Bayesian model averaging
  - very common
  - based on prior probabilities for potential models and priors for parameters.
  - Hoeting et al. (1999, Stat Science)
- Frequentist model averaging
  - Hjort and Claeskens (2003, JASA)
  - Yuan and Yang (2005, JASA)
  - Leung and Barron (2006, IEEE Info Th.)
  - Hansen (2007, Econometrica)

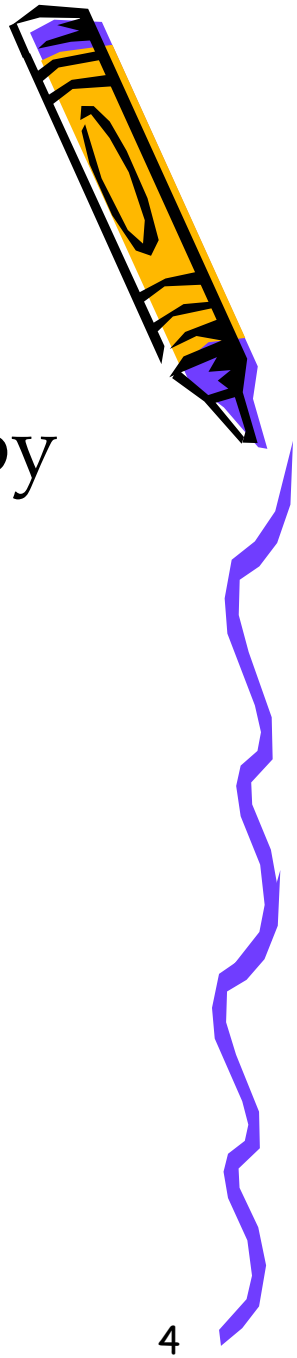


- Current paper motivated by Hansen (2007, Econometrica)

- Hansen's approach: Weights chosen by minimizing the Mallows criterion, equivalent to squared error in large samples.

- Model framework:

$$\underset{(n\times 1)}{y} = \underset{(n\times P)(P\times 1)}{H\theta} + \underset{(n\times 1)}{\varepsilon} \quad ; \quad \varepsilon \sim i.i.d.\left(0, \sigma^2\right)$$



Hansen's approach:

- Order the regressors at the outset, $X_1, X_2, X_3, \ldots X_p$
- Estimate a set of nested models:

$$y = X_1\theta_1 + \varepsilon_1 \;;$$

$$y = X_1\theta_1 + X_2\theta_2 + \varepsilon_2 \;;$$

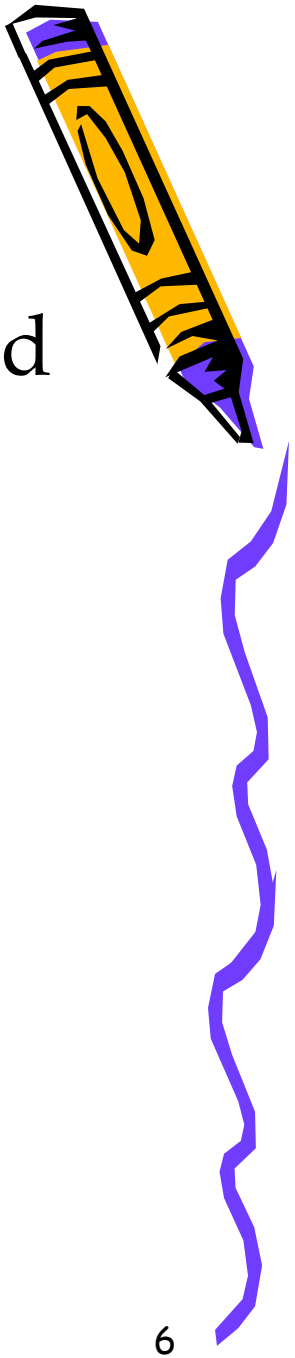
$$\vdots$$

$$y = X_1\theta_1 + X_2\theta_2 + \cdots + X_p\theta_p + \varepsilon$$

- Let H_p be an $n \times p$ ($\leq P$) matrix comprising the 1st p columns of H and ω_p is the weight.

- Hansen's (MMA) estimator :

$$\hat{\Theta}_m = \sum_{p=1}^{P} \omega_p \begin{pmatrix} (H_p' H_p)^{-1} H_p' y \\ 0 \end{pmatrix}$$



- Mallows criterion:

$$C(\omega) = \left(y - H\hat{\theta}\right)'\left(y - H\hat{\theta}\right) + 2\sigma^2 k(\omega)$$

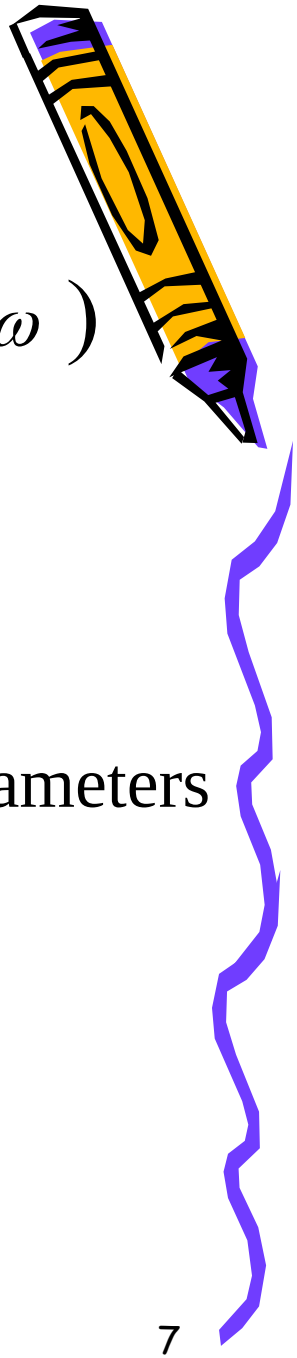
where

$$\omega = (\omega_1, \omega_2 \ldots \omega_P)'$$

and

$k(\omega)$ is the effective number of parameters

- $\hat{\omega} = \arg\min C(\omega)$



Difficulties with Hansen's (2007) approach:

1) explicit ordering of regressors

2) Estimation of nested models

$$y = X_1\theta_1 + \varepsilon_1 \;;$$
$$y = X_1\theta_1 + X_2\theta_2 + \varepsilon_2 \;;$$
$$\vdots$$
$$y = X_1\theta_1 + X_2\theta_2 + \cdots + X_p\theta_p + \varepsilon$$

(cannot handle, for example, combining (X_1, X_4, X_8) and (X_1, X_5, X_7)

3) criterion based on asymptotic justification

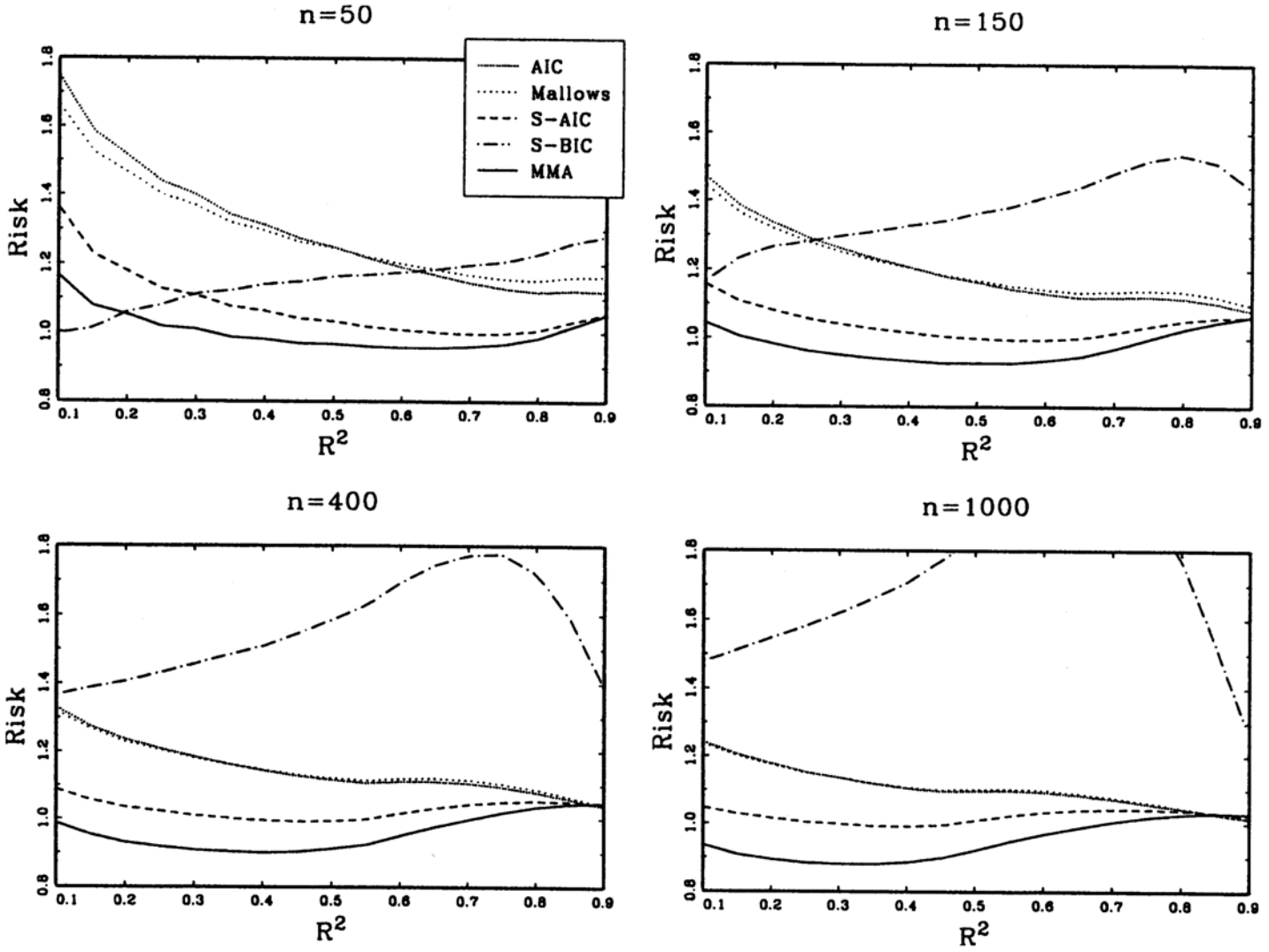


FIGURE 1.—$\alpha = 0.5$.

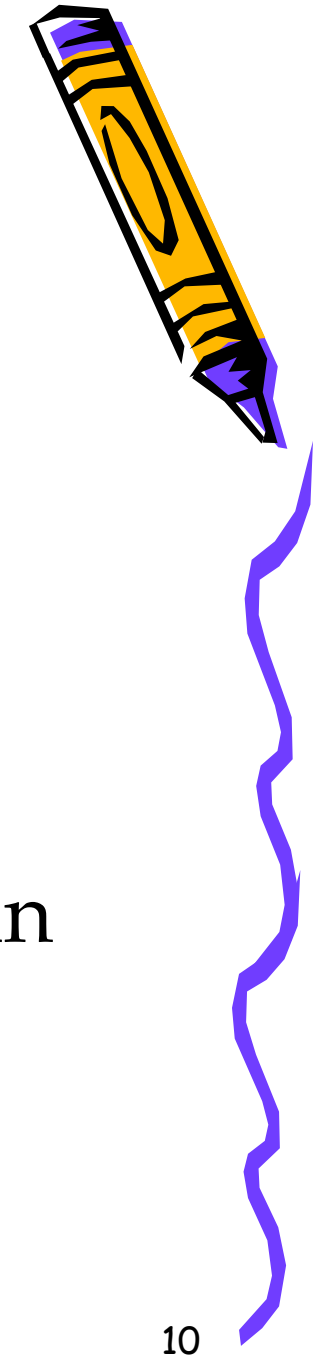
Alternative approach:

$$\underset{(n\times 1)}{y} = \underset{(n\times k)(k\times 1)}{X\beta} + \underset{(n\times m)(m\times 1)}{Z\gamma} + \underset{(n\times 1)}{\varepsilon} \quad ;$$

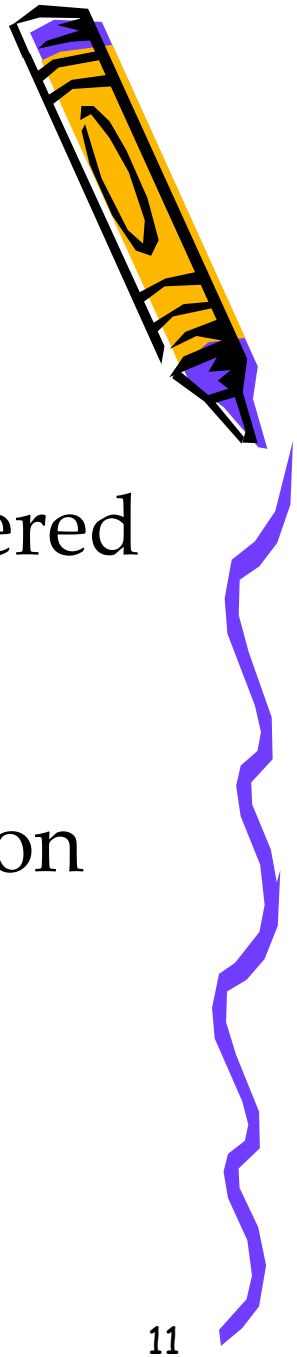
X : focus (required) regressors

Z : auxillary regressors

Framework follows Magnus and Durbin (1999, Econometrica)



Choice of weights



- when $m = 1$, Magnus (2002, Econometrics Journal) and Danilov (2005, Econometrics Journal) considered weight based on Laplace prior.

- Our approach: select weights based on the *MSE* of the weighted average estimator.

- With m auxiliary regressors in Z, there are 2^m models

- Unrestricted estimators :

$$b_u = b_r - Q\hat{\theta} \quad ; \quad \hat{\gamma}_u = D^2 Z'My$$

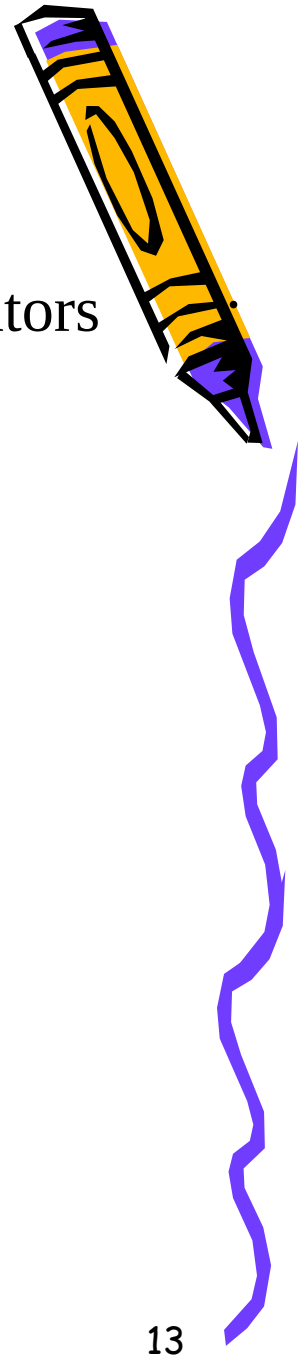
- Fully restricted estimators:

$$b_r = (X'X)^{-1}X'y \quad ; \quad \hat{\gamma}_r = 0$$

where $Q = (X'X)^{-1}X'ZD$, $D = (Z'MZ)^{-1/2}$,

$M = I_n - X(X'X)^{-1}X'$ and $\hat{\theta} = DZ'My \sim N(\theta, \sigma^2 I_m)$

with $\theta = D^{-1}\gamma$.



The i^{th} $\left(0 \le i \le 2^m\right)$ partially restricted estimators:

$$b_{(i)} = b_r - QW_i\hat{\theta}\ ; \qquad \hat{\gamma}_{(i)} = DW_i\hat{\theta}$$

where $W_i = I_m - P_i,\ P_i = DS_i\left\{S_i'D^2S_i\right\}^{-1}S_i'D$,

and S_i is a selection matrix of rank r_i.

- Traditional model selection chooses the "best" among the 2^m models.

- Frequentist Model Average estimators:

$$b_f = \sum_{i=1}^{2^m} \lambda_i b_{(i)} \quad ; \quad \hat{\gamma}_f = \sum_{i=1}^{2^m} \lambda_i \hat{\gamma}_{(i)}$$

  where $\lambda_i \geq 0$ and $\sum_{i=1}^{2^m} \lambda_i = 1.$

- Consider weights $\lambda_i = \lambda_i(\hat{\theta}, \hat{\sigma}^2)$.

- Write $W = \sum_{i=1}^{2^m} \lambda_i W_i$

Theorem 3.1

$$\hat{MSE}(b_f) = \hat{\sigma}^2 (X'X)^{-1} - \hat{\sigma}^2 QQ' + \{Q(I_m - W)\hat{\theta}\}^2 + \Psi(\hat{\theta}, \hat{\sigma}) + \{\psi(\hat{\theta}, \hat{\sigma}^2)\}'$$

where

$$\Psi(\hat{\theta}, \hat{\sigma}^2) = ((n-k-m)/2)(\hat{\sigma}^2)^{-(n-k-m)/2+1} \int_0^{\hat{\sigma}^2} t^{(n-k-m)/2-1} \Psi_1(\hat{\theta}, t)\, dt$$

and

$$\Psi_1(\hat{\theta}, t) = Q\left\{W + \sum_{i=1}^{2^m} \left(\partial \lambda_i(\hat{\theta}, t)/\partial \hat{\theta}\right) \hat{\theta}' W_i\right\} Q'$$

$$\hat{R}(b_f) = \hat{\sigma}^2 tr(X'X)^{-1} - \hat{\sigma}^2 tr(QQ') + \hat{\theta}(I_m - W)Q'Q(I_m - W)\hat{\theta}$$
$$+ 2tr\{\Psi(\hat{\theta}, \hat{\sigma}^2)\}$$

One problem with minimizing $\hat{R}(b_f)$ is that

$$\Psi(\hat{\theta}, \hat{\sigma}^2) = ((n-k-m)/2)(\hat{\sigma}^2)^{-(n-k-m)/2+1} \int_0^{\hat{\sigma}^2} t^{(n-k-m)/2-1} \Psi_1(\hat{\theta}, t)dt$$

is complex.

Solution :

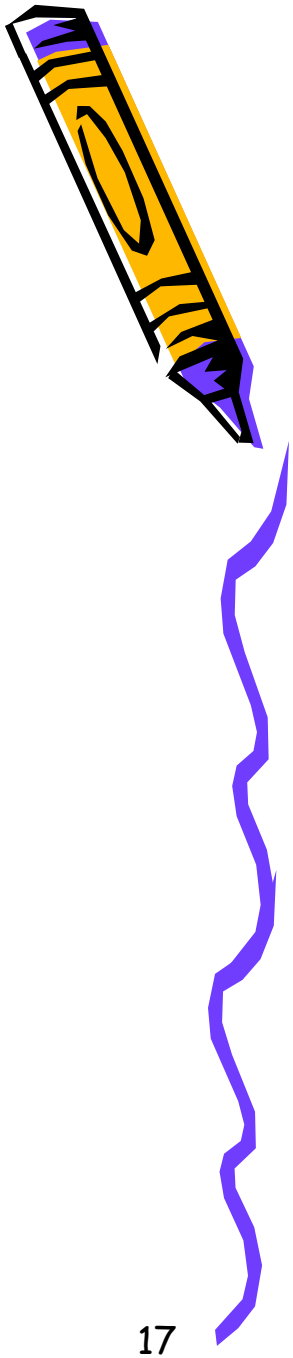
Replace $\Psi\left(\hat{\theta}, \hat{\sigma}^2\right)$ by

$\hat{\sigma}^2 \Psi_1\left(\hat{\theta}, \hat{\sigma}^2\right)$

where

$$\Psi_1(\hat{\theta}, t) = Q\left\{W + \sum_{i=1}^{2^m}\left(\partial\lambda_i(\hat{\theta}, t)/\partial\hat{\theta}\right)\hat{\theta}' W_i\right\} Q'.$$

since $E\left\{\psi\left(\hat{\theta}, \hat{\sigma}^2\right)\right\} = \sigma^2 E\left\{\Psi_1\left(\hat{\theta}, \hat{\sigma}^2\right)\right\}$.



So, we have

$$\hat{R}_a(b_f) = \hat{\sigma}^2 tr(X'X)^{-1} - \hat{\sigma}^2 tr(QQ') + \hat{\theta}(I_m - W)Q'Q(I_m - W)\hat{\theta}$$
$$+2tr\{\Psi(\hat{\theta}, \hat{\sigma}^2)\},$$

Write $\lambda_i(\hat{\theta}, \hat{\sigma}^2) = \left(a_i(\hat{\sigma}_i^2)^c\right) / \left(\sum_{i=1}^{2^m} a_i(\hat{\sigma}_i^2)^c\right)$,

where a_i's are positive constants and c is a non-positive constant.

S-AIC (Buckland et al. (1997, Biometrics)):

$$a_i = \exp\{-(q_i + 1)\} \; ; \; c = -n/2$$

S-BIC : $a_i = -n^{-(q_i+1)/2}$; $c = -n/2$

S-AICC (Hurvich and Tsai (1989, Biometrika)):

$$a_i = \exp\{-n(q_i+1)/(n - q_i - 2)\} \quad ; \quad c = -n/2$$

Recall that

$$\Psi_1\left(\hat{\theta},\hat{\sigma}^2\right)=Q\left\{W+\sum_{i=1}^{2^m}\left(\partial\lambda_i\left(\hat{\theta},\hat{\sigma}^2\right)/\partial\hat{\theta}\right)\hat{\theta}'W_i\right\}Q'\ .$$

Now,

$$\partial\lambda_i\left(\hat{\theta},\hat{\sigma}^2\right)/\partial\hat{\theta}=(2/n)c\lambda_i\left(\hat{\theta},\hat{\sigma}^2\right)\left\{\left(\hat{\sigma}_i^2\right)^{-1}(I_m-W_i)-\sum_{i=1}^{2^m}\lambda_i\left(\hat{\theta},\hat{\sigma}^2\right)\left(\hat{\sigma}_i^2\right)^{-1}\right\}$$

$$\times(I_m-W_i)\}\hat{\theta}, \qquad (*)$$

Putting (*) in $\psi_1\left(\hat{\theta},\hat{\sigma}^2\right)$, we have

$$\hat{R}_a\left(b_f\right)=\hat{\sigma}^2 tr\left(X'X\right)^{-1}-\hat{\sigma}^2 tr\left(QQ'\right)+\lambda' L\lambda-\left(4/n\right)c\hat{\sigma}^2\lambda G\lambda$$
$$+2\hat{\sigma}^2\lambda'\phi+\left(4/n\right)c\hat{\sigma}^2\lambda' g,$$

where

$L=(l_{ij})$, $G=(g_{ij})$,

$$l_{ij}=\hat{\theta}'\left(I_m-W_i\right)Q'Q\left(I_m-W_j\right)\hat{\theta}\ ,$$

$$g_{ij}=\left(\hat{\sigma}_j^2\right)^{-1}\hat{\theta}'W_iQ'Q\left(I_m-W_j\right)\hat{\theta},\ \ i,j=1,\ldots 2^m,$$

and ϕ each be a $2^m\times 1$ vector with g consisting of the diagonal elemtns of G and the ith element of ϕ be $tr\left(QW_iQ'\right),\ \ i=1,\ldots 2^m.$

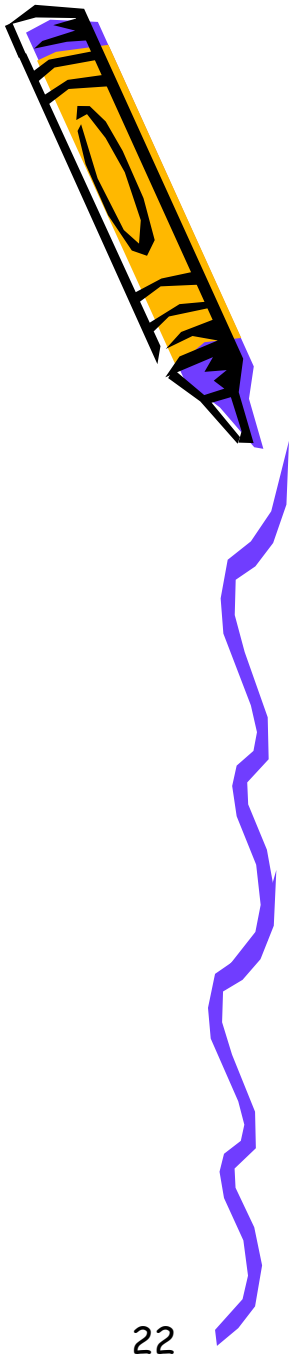
Interesting special case

Setting $c = 0$ and considering only mixing b_u and b_r, then minimization criterion leads to

$$b_{js} = \left\{1 - \frac{\hat{\sigma}^2 tr(Q'Q)}{\| b_u - b_r \|^2}\right\} b_u + \frac{\hat{\sigma}^2 tr(Q'Q)}{\| b_u - b_r \|^2} b_r ,$$

i.e. James and Stein estimator !!

Optimal predictor



Let $\hat{\mu}_f = H\,\hat{\theta}_f$,

where $\hat{\theta}_f = \left(b'_f\ ,\ \hat{\gamma}'_f\right)$

and $\hat{\gamma}_f = \sum_{i=1}^{2^m} \lambda_i\left(\hat{\theta}\ ,\ \hat{\sigma}^2\right)\hat{\gamma}_{(i)}$ is the

estimator of γ corresponding to b_f

$$\hat{MSE}(\hat{\mu}_f) = \hat{\sigma}^2 X(X'X)^{-1}X' - \varphi\left(\hat{\theta}, \hat{\sigma}^2, XQ, (XQ)'\right) - \varphi\left(\hat{\theta}, \hat{\sigma}^2, XQ, (ZQ)'\right)$$
$$- \varphi\left(\hat{\theta}, \hat{\sigma}^2, ZD, (XQ)'\right) + \varphi\left(\hat{\theta}, \hat{\sigma}^2, ZD, (ZD)'\right)$$

where

$$\varphi(\hat{\theta}, \hat{\sigma}^2, C_1, C_2) = -\hat{\sigma}^2 C_1 C_2 + C_1\left\{(I_m - W)\hat{\theta}\right\}^{\otimes 2} C_2 + C_1 \Xi(\hat{\theta}, \hat{\sigma}^2) C_2 + C_1 \{\Xi(\hat{\theta}, \hat{\sigma}^2)\}' C_2,$$

$$\Xi(\hat{\theta}, \hat{\sigma}^2) = \frac{n-k-m}{2} (\hat{\sigma}^2)^{-\frac{n-k-m}{2}+1} \int_0^{\hat{\sigma}^2} t^{\frac{n-k-m}{2}-1} \Xi_1(\hat{\theta}, t) dt$$

and

$$\Xi_1(\hat{\theta}, t) = W + \sum_{i=1}^{2^m} \frac{\partial \lambda_i(\hat{\theta}, t)}{\partial \hat{\theta}} \hat{\theta}' W_i$$

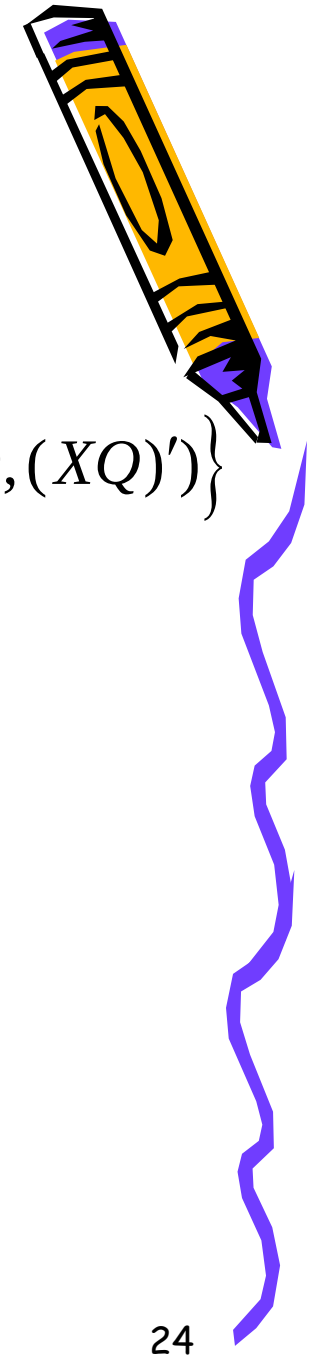
The trace of $MSE(\hat{\mu}_f)$ is

$$\hat{R}(\hat{\mu}_f) = k\hat{\sigma}^2 + tr\left\{\varphi(\hat{\theta}, \hat{\sigma}^2, XQ, (XQ)')\right\} - 2tr\left\{\varphi(\hat{\theta}, \hat{\sigma}^2, ZD, (XQ)')\right\}$$

which is approximately equal to

$$\hat{R}_a(\hat{\mu}_f) = (k-m)\hat{\sigma}^2 + \lambda' \bar{L} \lambda + 2\hat{\sigma}^2 \lambda' \bar{\phi} - (4/n)c\hat{\sigma}^2 \lambda' \bar{G} \lambda$$

by analogy of previous analysis.



Experiment 1

$$y_i = \sum_{j=1}^{7} \theta_j x_{ji} + e_i ,$$

$$x_{1i} = 1,\ x_{2i} \sim N(0,1),\ x_{3i} \sim N(0,1)$$

$$x_{4i} \sim N(4,1),\ x_{5i} \sim u(0,1).$$

$$x_{6i} \sim N(4,1),\ x_{7i} \sim U(0,1),\ e_i \sim N(0,1);$$

$$\theta_j = c_{1j}^{-\alpha} ;$$

$$R^2 = c_1^2(\tau' v)/(1 + c_1^2(\tau' v)) ;$$

v is a vector comprising the variances of the regressors ;

$$\tau_j = j^{-2\alpha}$$

$$n = 30 , 80 , 150 , 300$$

$$\alpha = 0.5, 1.0, 1.5$$

OPT (Optimal Frequentist MA) estimator

Focus regressors : x_1, x_2, x_3, x_4

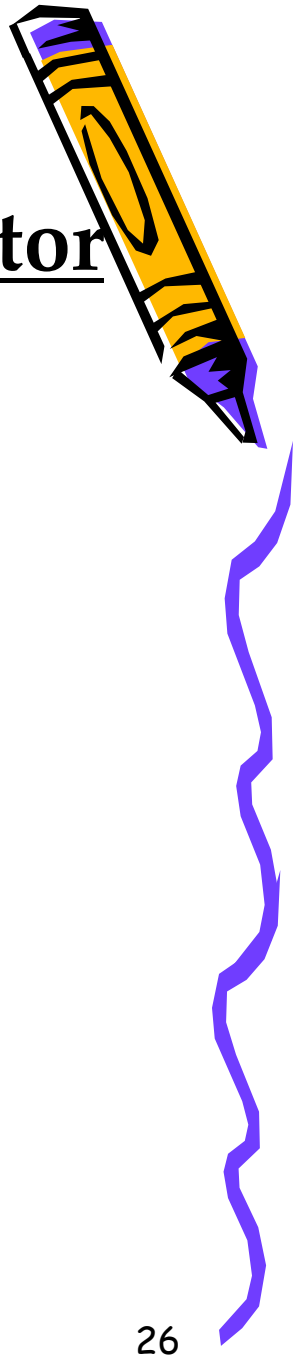
Auxiliary regressors : x_5, x_6, x_7

(i.e., 2^3 models)

MMA (Hansen's) estimator

Order regressors $x_1, x_2, x_3, x_4, x_5, x_6, x_7$

(i.e., 7 models)



$$L_1 = \left[\left(\hat{y} - \sum_{j=1}^{7} \theta_j x_j \right)' \left(\hat{y} - \sum_{j=1}^{7} \theta_j x_j \right) \right]$$

$$L_2 = \left[\sum_{j=1}^{4} \left(\hat{\theta}_j - \theta_j \right)^2 \right]$$

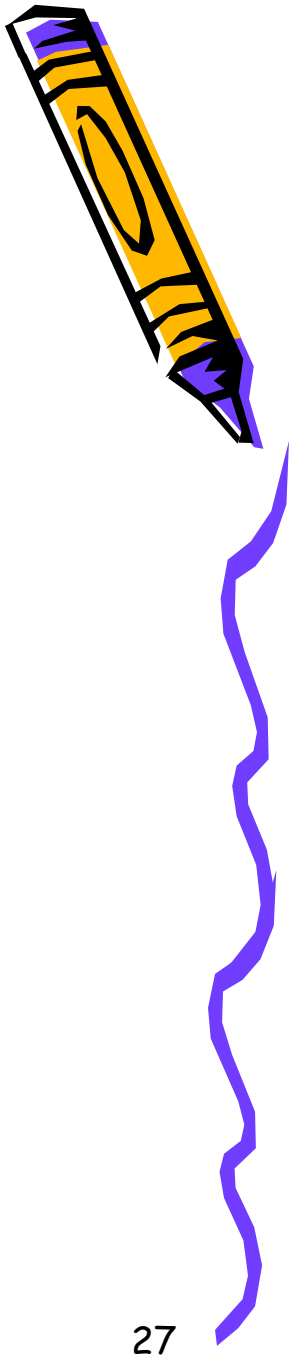


Figure 1: Risk under loss L_1 ($\alpha = 0.5$)

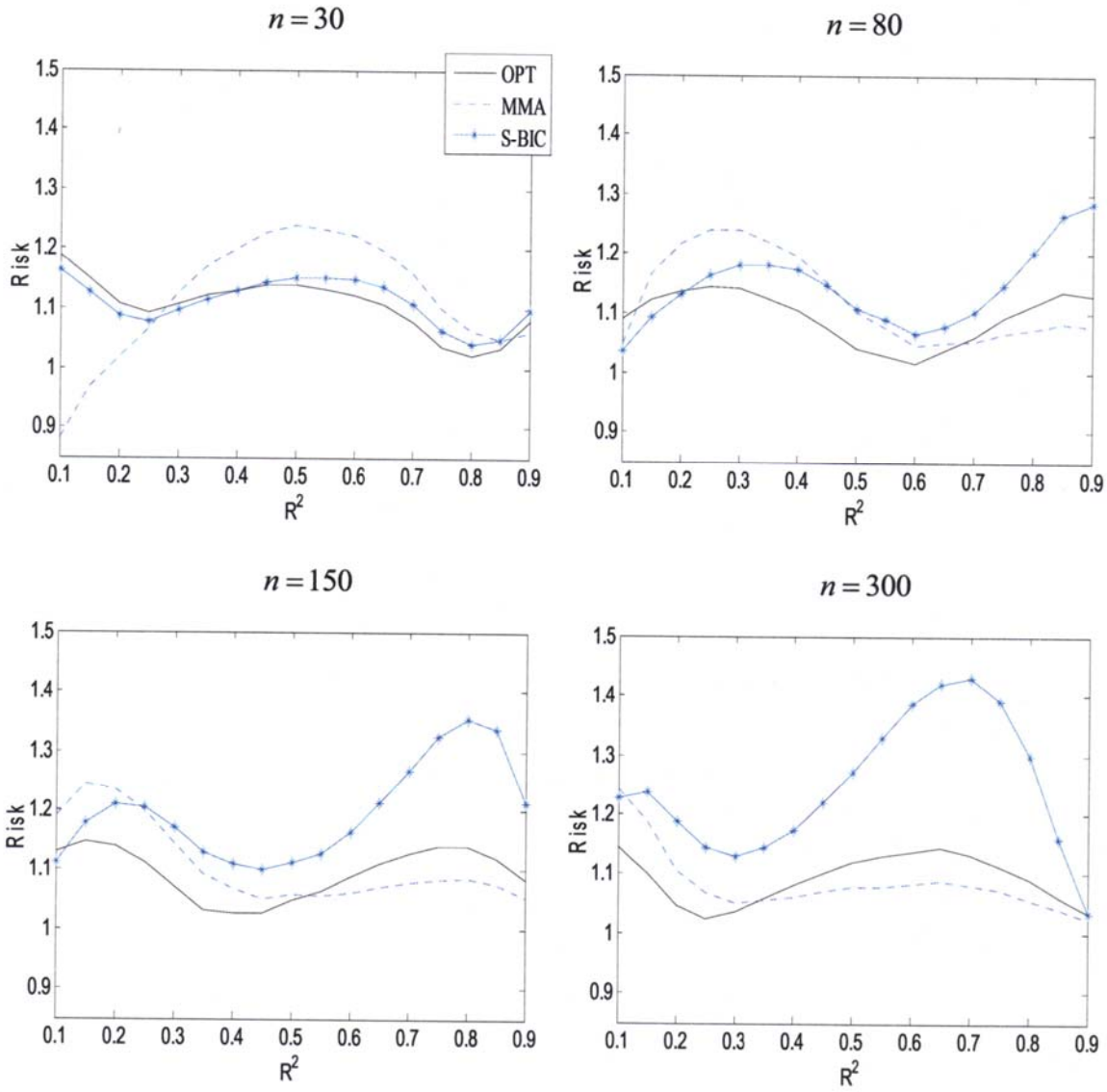


Figure 2: Risk under loss L_1 ($\alpha = 1.0$)

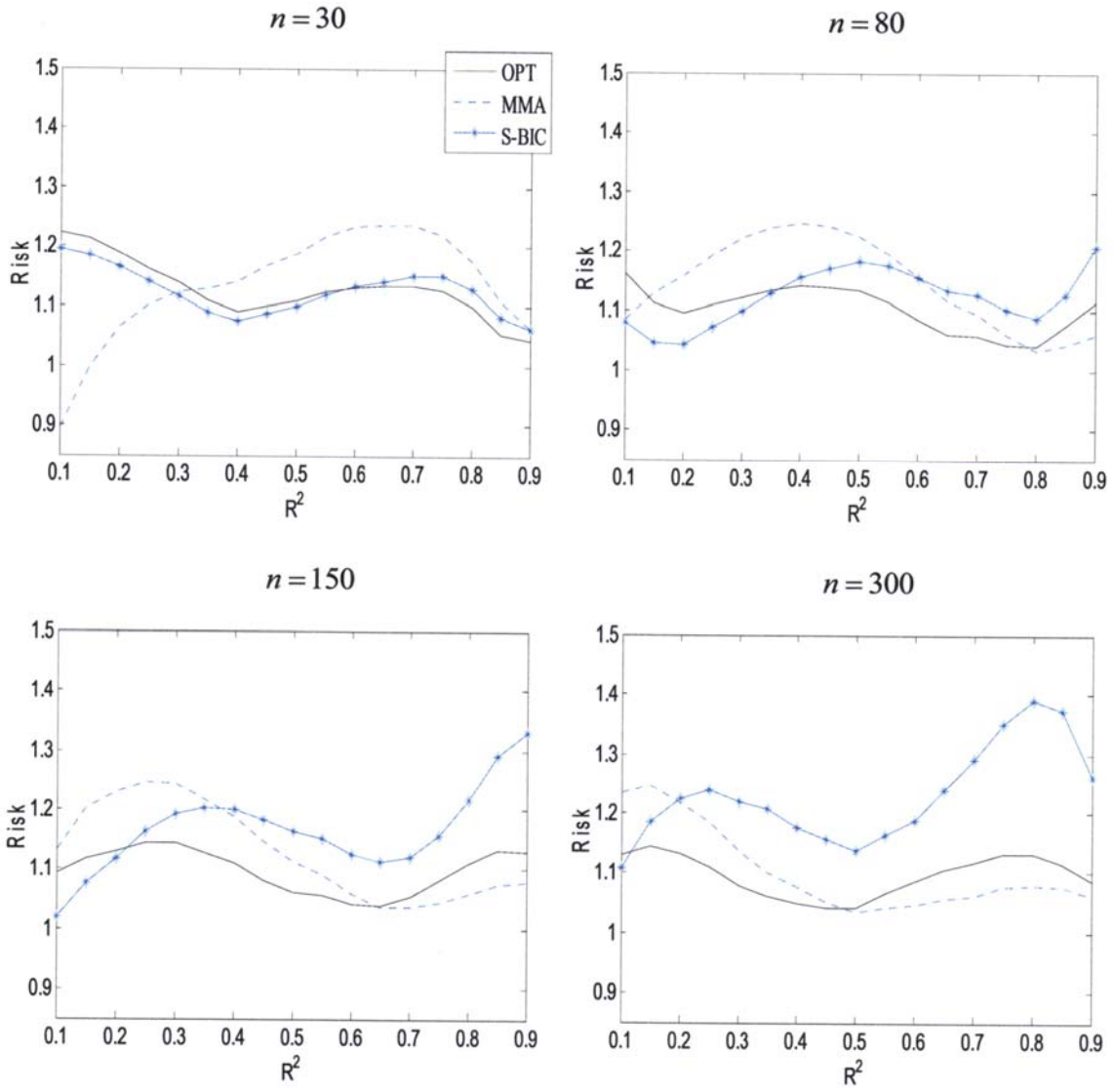


Figure 3: Risk under loss L_1 ($\alpha = 1.5$)

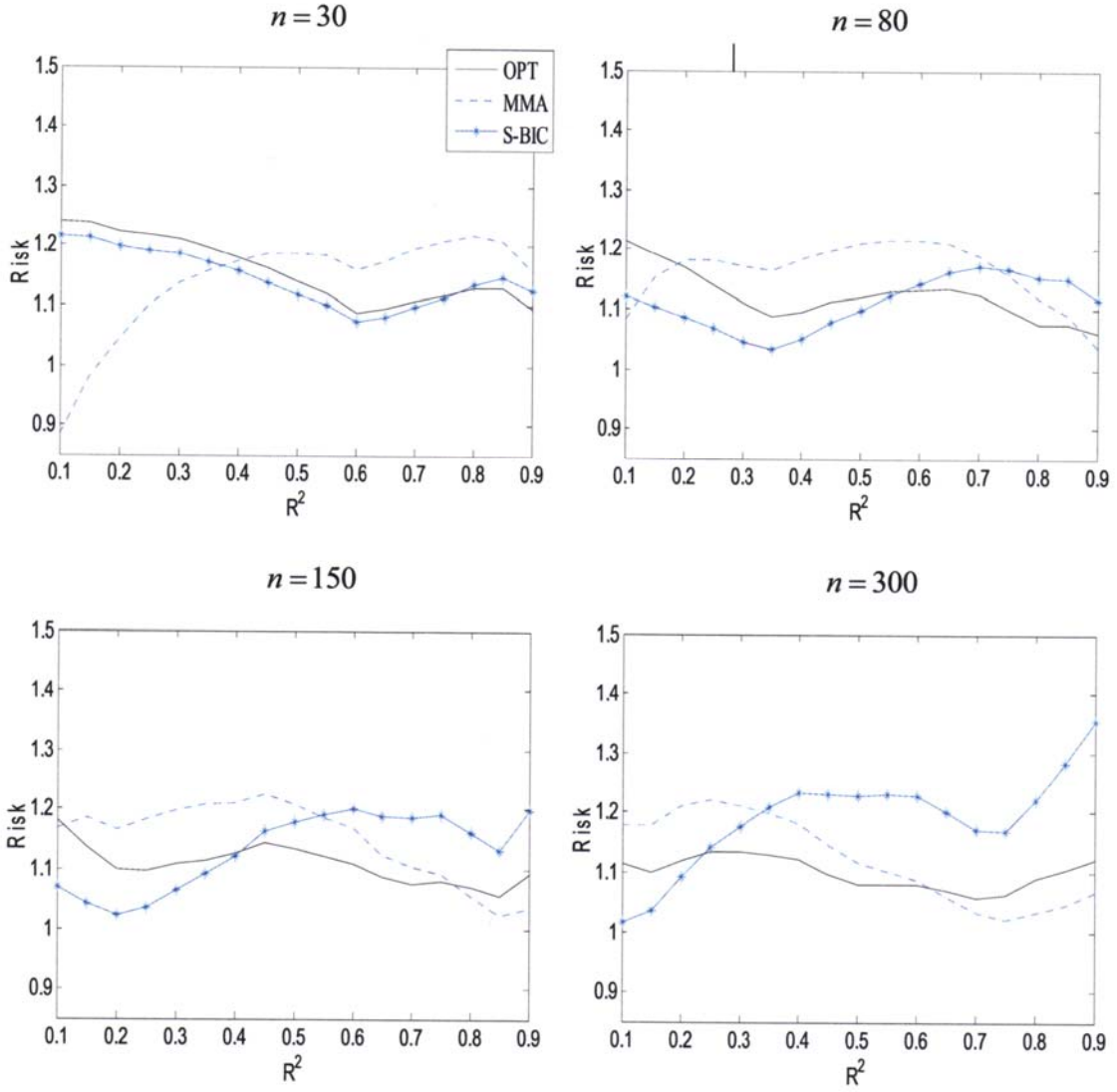
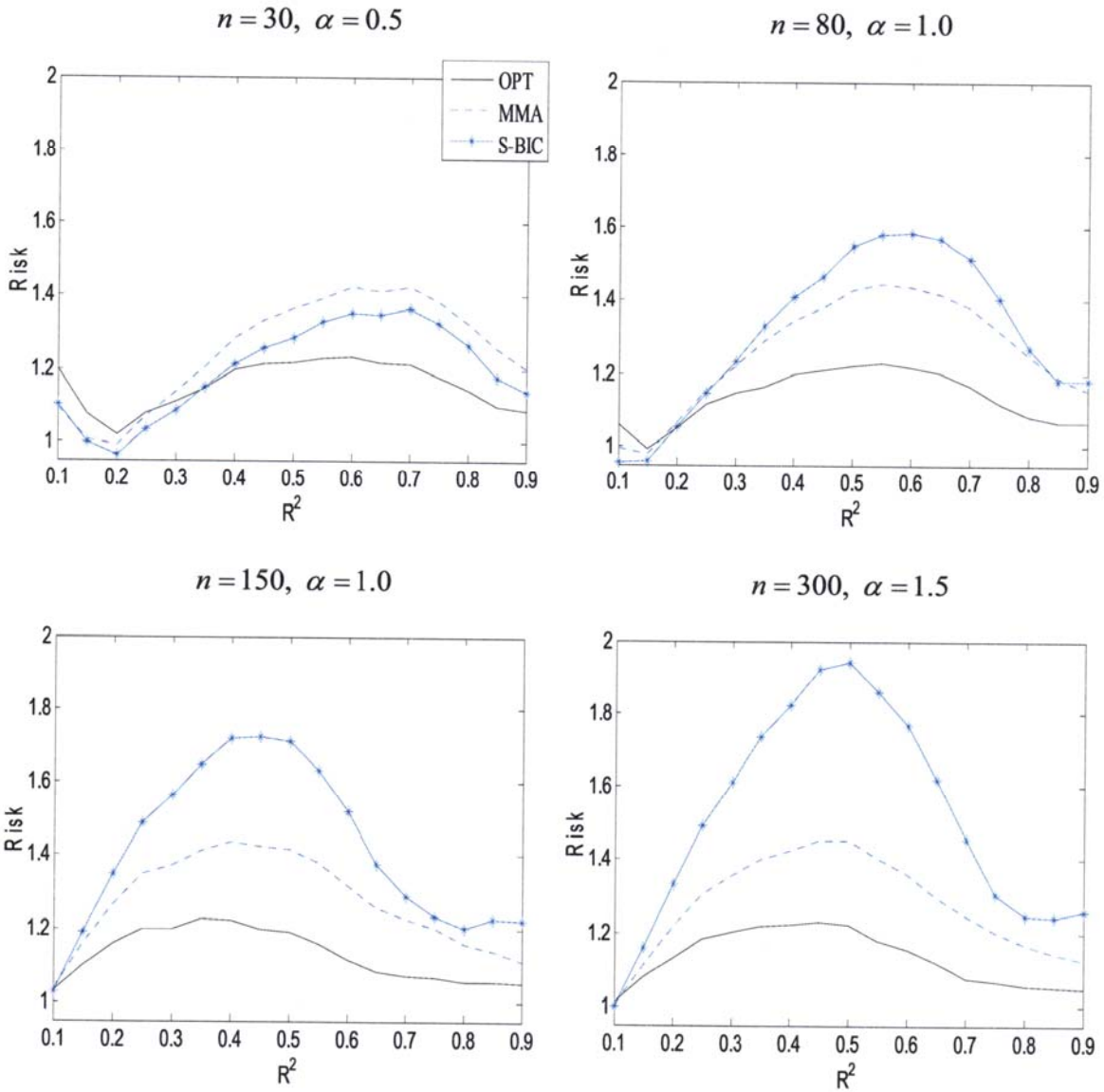
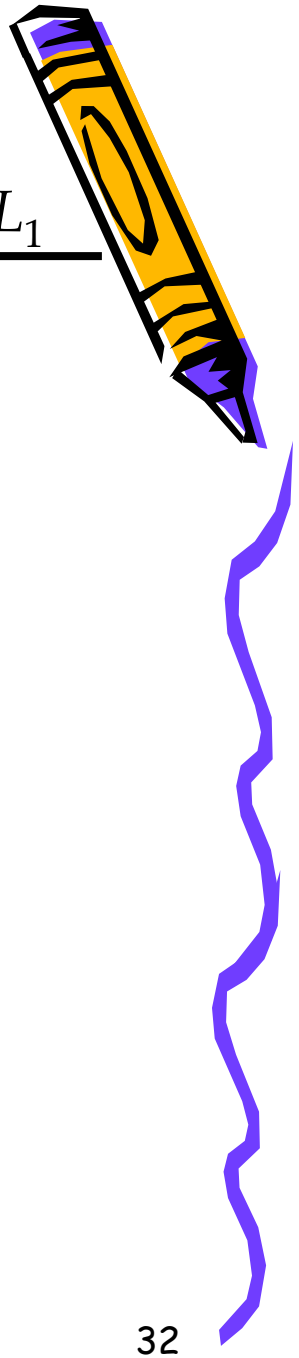


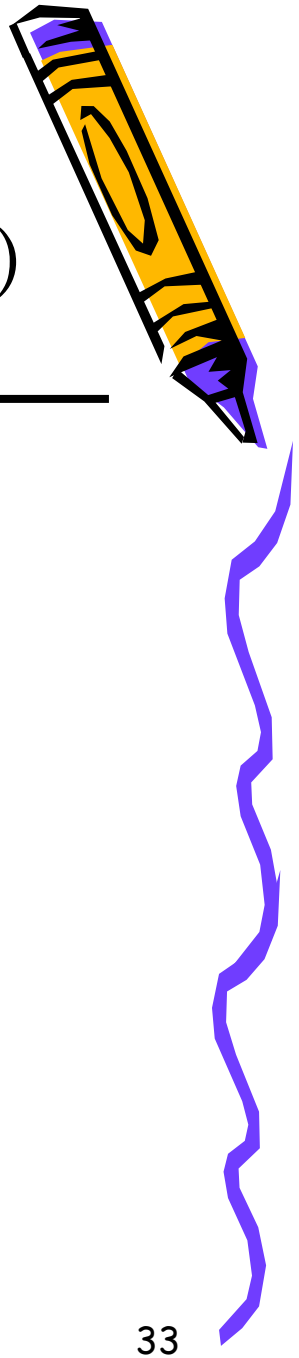
Figure 4: Risk under loss L_2



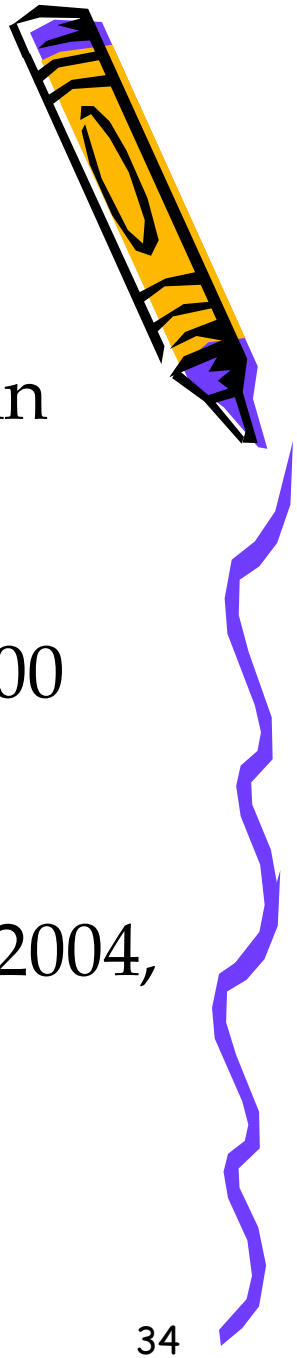
α	n	Risk (OPT) < Risk (MMA) under L_1
0.5	30	73.99%
	80	69.77%
	150	53.34%
	300	30.88%
1.0	30	72.47%
	80	80.36%
	150	68.45%
	300	47.25%
1.5	30	61.51%
	80	84.98%
	150	82.98%
	300	64.95%



α	n	Risk (OPT) < Risk (MMA) under L_2
0.5	30	79.96%
1.0	80	91.13%
1.0	150	100%
1.5	300	98.67%



Experiment 2



- Data taken from Pearson and Timmermann (1994, J. of Forecasting)
- Predictability of excess returns for S & P 500 Index
- Same data used by Danilov and Magnus (2004, J. of Forecasting)

$$y_t = \beta_1 + \beta_2 PI_{t-2} + \beta_3 DI3_{t-1} + \beta_4 SPREAD_{t-1} + \gamma_1 YSP_{t-1} + \gamma_2 DIP_{t-1} + \gamma_3 PER_{t-1} + \gamma_4 DLEAD_{t-2} + \varepsilon_t$$

and definition:

y_t = excess returns,

PI_{t-2} = annual inflation rate (lagged two periods),

$DI3_{t-1}$ = change in 3-month T-bill rate (lagged one period),

$SPREAD_{t-1}$ = credit spread (lagged one period),

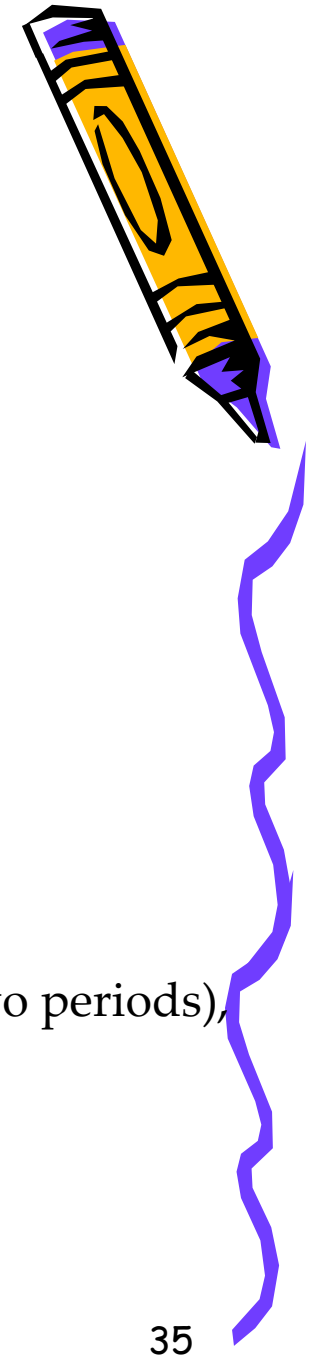
YSP_{t-1} = dividend yield on SP500 portfolio (lagged one period),

DIP_{t-1} = annual change in industrial production (lagged one period),

PER_{t-1} = price-earnings ratio (lagged one period),

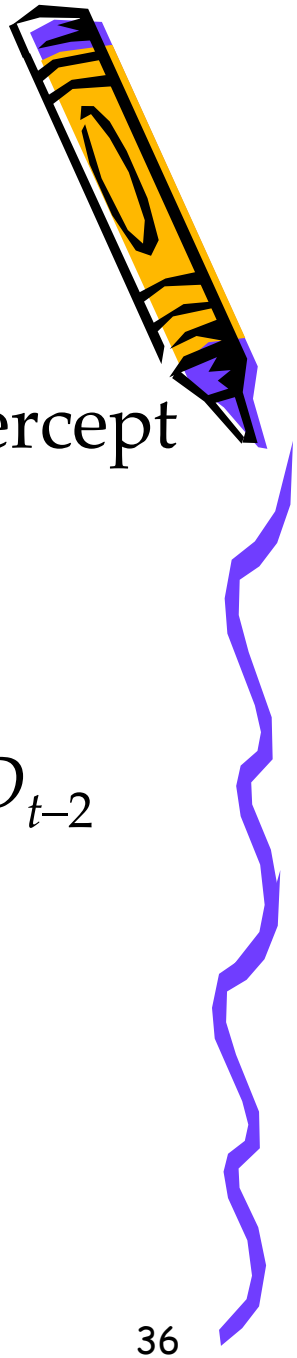
$DLEAD_{t-2}$ = annual change in leading business cycle indicator (lagged two periods),

t = 1956 – 2001 (46 observations).



Focus regressors : Pi_{t-2}, $DI3_{t-1}$, SPREAD$_{t-1}$ and intercept

Auxiliary regressors : YSP_{t-1}, DIP_{t-1}, PER_{t-1} and $DLEAD_{t-2}$

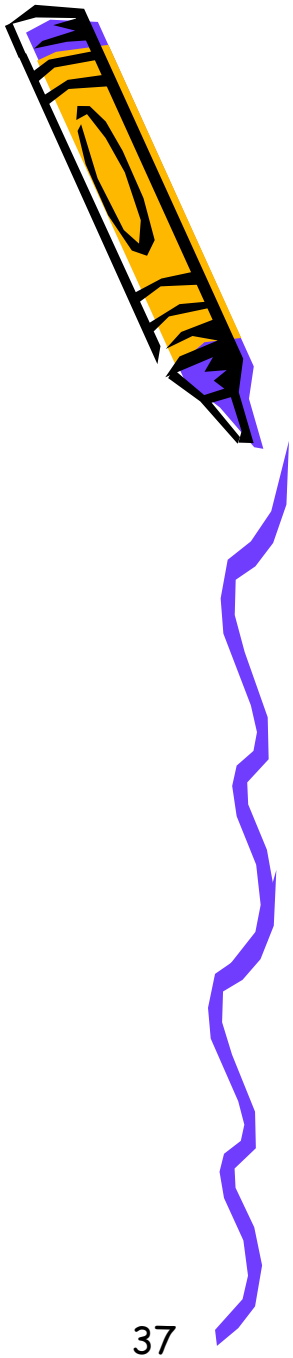


OPT estimator

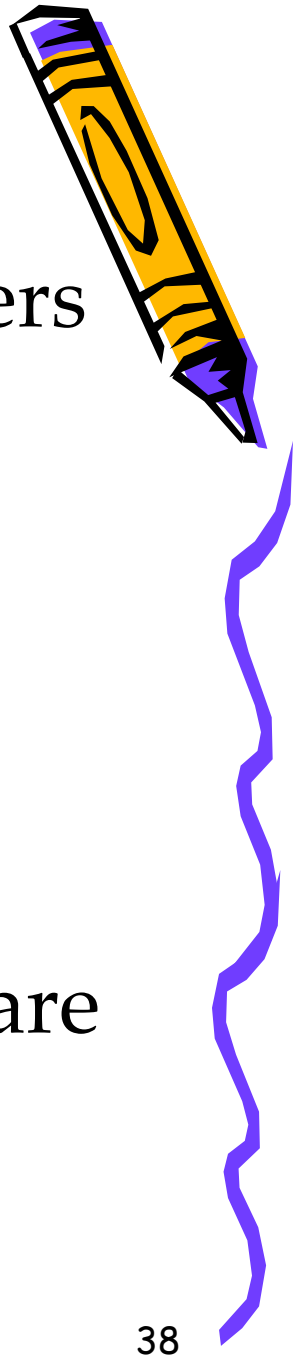
$2^4 = 16$ models

MMA estimator

8! = 40320 possible ordering sequences

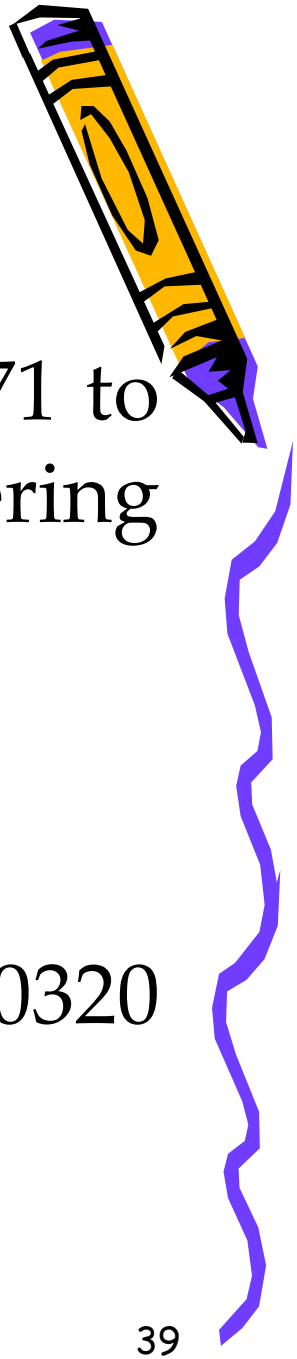


Design of Monte Carlo Study

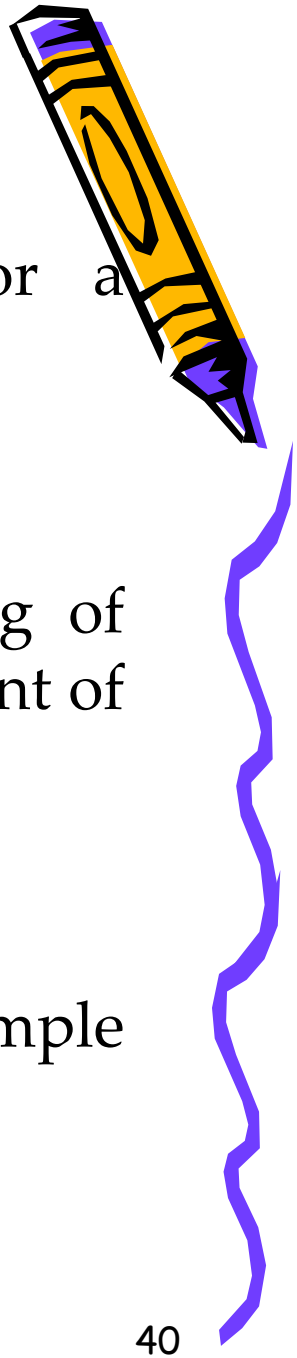


- Uses OLS estimates as true parameters
- y_t in each round of simulation is obtained by drawing 46 random disturbances with replacement from OLS residuals.
- A total of 100 Monte-Carlo samples are drawn.

- $R(OPT)$ under L_1 is 0.0878
- $R(MMA)$ under L_1 ranges from 0.0771 to 0.1057 depending on the ordering pattern.
- Average $R(MMA) = 0.0942$
- $R(OPT) < R(MMA)$ in 35778 out of 40320 or 88.7% of cases.



Conclusions



- Alternative way to select model weights for a frequentist model average estimator

- Merits
  - framework does not require explicit ordering of regressors (ordering pattern is a key determinant of Hansen's estimator)
  - Finite sample justifications

- Work ahead : extend present analysis to out-of-sample forecasts.