淺談統計模型之選取

段慶剛

中研院統計所

一〇五、八、十三

(1)

- 模型 (Model): 用來描述自然或社會科學資料的數學函數 ($y = f_{\underset{\sim}{\beta}}(\underset{\sim}{x})$).

  $\underset{\sim}{\beta}$: 未知參數　　$\underset{\sim}{x}$: 解釋變數

- 統計模型 (Statistical Model).

$$y = f_{\underset{\sim}{\beta}}(\underset{\sim}{x}) + \varepsilon$$

ε: 隨機誤差

例如:

1. 線性迴歸模型:

$$y = \underset{\sim}{x}'\underset{\sim}{\beta} + \varepsilon \sim \begin{array}{l} E(\varepsilon) = 0 \\ Var(\varepsilon) = \sigma^2 \end{array}$$

$\underset{\sim}{\beta}$: 未知參數　　σ^2: 未知

2. 非線性迴歸模型：

$$Y = x_1 x_2^{\beta_1} x_3^{\beta_2} + \varepsilon$$

↓

Cobb-Douglas production function

3. 時間序列模型：

$$Y_t = a Y_{t-1} + \varepsilon_t$$ （一階自我迴歸模型）

first-order autoregressive model AR(1) model

$$Y_t = a_1 Y_{t-1} + \cdots + a_p Y_{t-p} + \varepsilon_t$$ AR(p) model

$$Y_t = a_1 Y_{t-1} + \cdots + a_p Y_{t-p} + \beta_1 X_{t1} + \cdots + \beta_q X_{tq} + \varepsilon_t$$

內生變數　　外生變數

ARX model (autoregressive exogenous model)

(3)

Q: 當有很多候選模型可供選擇時，吾人應採何種作為？

A: 模型選取（Model Selection）

Q: 選模的目的為何？

A: 1. 解釋現象

2. 預測

④

⊙ 选模方法简介

A. Testing-based methods.

ⓐ Forward Selection

ⓑ Backward Elimination

ⓒ Stepwise procedure

B. prediction-based methods.

ⓐ Mallows' C_p

(b) Cross-validation (CV)

(i) Delete-1-out CV

(ii) Delete-k-out CV

(c) Accumulated Prediction Error (APE).

(i) original APE

(ii) $APE_{[n]}$

C. Information-based methods

(a) Akaike's Information Criterion (AIC)

(b) Bayesian Information Criterion (BIC)

(c) Hannan-Quinn Information Criterion (HQ)

(6)

① 选模方法的表现.

AIC
C_p
Delete-1-out CV
APE_{δ_n}
Mallows' C_p

→ 预测表现良好

BIC
HQ
Delete-n_v-out CV ($n_v \to \infty$)
Original APE

→ 利于解释现象.

⑦

⊙ 逐模方法的模擬表現

⊙ 限制：AIC與BIC的戰爭.

⑧ ① 高維資料的模型選取

A. 向後淘汰

B. LASSO (Least Absolute Shrinkage and Selection Operator)

C. Greedy Algorithm

a. pure Greedy Algorithm (pGA)

b. Orthogonal Greedy Algorithm (OGA)

D. 傳統與創新：供應方法的高低

低手

OGA + HDIC + Trim

↓ Forward selection

↓ Information criteria

↓ Backward elimination

E. 實例分析：晶圓資料分析之由共演

F. 高低供應方法之競爭未來.

END.

(a.1)

Forward Selection.

$$y_i = x_i'\beta + \varepsilon_i$$

$$= \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i$$

Step 1: 选取与 y 相关性最大的 x_i, $i=1\cdots p$.

Step 2: 计算 $F = \dfrac{R(x_2|x_1)}{s^2(x_1,x_2)}$

$$= \frac{\sum_{i=1}^{n}(y_i-\hat\beta_0-\hat\beta_1 x_{i1})^2 - \sum_{i=1}^{n}(y_i-\tilde\beta_0-\tilde\beta_1 x_{i1}-\tilde\beta_2 x_{i2})^2}{\frac{1}{n-3}\sum_{i=1}^{n}(y_i-\tilde\beta_0-\tilde\beta_1 x_{i1}-\tilde\beta_2 x_{i2})^2}$$

测量 x_2 的"边际"贡献大

测量"噪音"

Step 3. 设定 F_{IN}

Step 4. 若当前变量给出 F_{IN}，则停止加入新变量.

(a.2)

- Stepwise :

  設定 Fout

- Backward elimination

  從 Full model 往下刪減變數

  設定 Fout .

a.3

Mallows' C_p.

考虑

$$\underset{\sim}{Y}^{New} = X_p \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}^{New},$$

其中 $X_p = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$，$\underset{\sim}{\varepsilon}^{New}$ 为 $\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

为与原回归的随机误差.

新的观察值，代表"未来".

(a4)

· 衡量模型 J 的預測表現，其中

$J \subseteq \{1, \dots, p\}$:

$$E\left\{ \sum_{i=1}^{n} \left(y_i^{New} - \hat{\beta}_J' x_i(J) \right)^2 \right\}$$

$x_i(J)$ → 模型 J 在第 i 點對應的解釋向量

$\hat{\beta}_J$ → 模型 J 對應的最小平方估計，其使用的資料為 $y_1, \dots, y_n$ → 視為看得到的資料

估計

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_J x_i(J) \right)^2 + 2\, \#(J)\, \hat{\sigma}^2$$

$\#(J)$：參數個數

$\hat{\sigma}^2$：σ^2 的不偏估計：$\frac{1}{n-p-1} \sum_{i=1}^{n} \left(y_i - \hat{\beta} x_i \right)^2$

$2\,\#(J)\,\hat{\sigma}^2$ → 偏誤修正量 (Bias correction term)

(a5)

CV:

对于给定的 J:

$$CV_1(J) = \sum_{i=1}^{n} \left(y_i - \hat{\beta}'_{J,-i} \underset{\sim}{x}_i(J) \right)^2 \quad \longleftarrow \text{delete-1-out CV}$$

$$CV_{n_v}(J) = \sum_{\#(S)=n_v} \sum_{i \in S} \left(y_i - \hat{\beta}'_{J,-S} \underset{\sim}{x}_i(J) \right)^2$$

$S \subseteq \{1, \dots n\}$

$\longleftarrow$ delete-n_v-out CV

Training data (S^c): $\{ y_i, \underset{\sim}{x}_i \}$, $i \in S^c$

Testing data (S): $\{ y_i, \underset{\sim}{x}_i \}$, $i \in S$.

(a6)

APE (for autoregressions).

$$APE = \sum_{i=1}^{n-1} \left(y_{i+1} - \hat{\beta}_J' \, y_{i,J} \right)^2 , \quad y_{i,J} = \{ y_{i+1-j} ; j \in J \} .$$

$$APE_{\delta_n} = \sum_{i=n\delta_n}^{n-1} \left(y_{i+1} - \hat{\beta}_J' \, y_{i,J} \right)^2 , \quad \frac{1}{n} \le \delta_n \le 1 - \frac{1}{n} .$$

(a7)

AIC:

$$-2 \log f_{\hat{\underset{\sim}{\theta}}}(\underset{\sim}{y}) + 2\{\text{No. of parameters}\}.$$

BIC:

$$-2 \log f_{\hat{\underset{\sim}{\theta}}}(\underset{\sim}{y}) + \log^n \{\text{No. of parameters}\}.$$

HQ (Hannan & Quinn)

$$-2 \log f_{\hat{\underset{\sim}{\theta}}}(\underset{\sim}{y}) + c(\log \log^n)\{\text{No. of parameter}\},$$

$c > 2$.

(a8) Reduced to the linear case:

$$AIC: \quad \log \hat{\sigma}_J^2 + \frac{2}{n} \#(J)$$

$$BIC: \quad \log \hat{\sigma}_J^2 + \frac{\log n}{n} \#(J)$$

$$HQ: \quad \log \hat{\sigma}_J^2 + \frac{c \log\log n}{n} \#(J)$$

In general,

$$IC_{p_n}: \quad \log \hat{\sigma}_J^2 + \frac{p_n}{n} \#(J).$$

(1)

Akaike's Information Criterion (AIC)

- Let $g(\underset{\sim}{y}^*) = g(y_1^*, \ldots, y_n^*)$ denote the likelihood function of $\underset{\sim}{y}^*$

  (true pdf of $\underset{\sim}{y}^*$)

  and $f_{\underset{\sim}{\theta}}(\underset{\sim}{y}^*)$ is a family of approximation models indexed by $\underset{\sim}{\theta}$.

  (A family of pdfs indexed by $\underset{\sim}{\theta}$.)

- The "distance" between $g(\underset{\sim}{y}^*)$ and $f_{\underset{\sim}{\theta}}(\underset{\sim}{y}^*)$ is measured by

  Kullback-Leibler distance

$$KL(f_{\underset{\sim}{\theta}}, g) = -E_g\left(\log \frac{f_{\underset{\sim}{\theta}}(\underset{\sim}{y}^*)}{g(\underset{\sim}{y}^*)}\right)$$

$$= -E_g\left(\log f_{\underset{\sim}{\theta}}(\underset{\sim}{y}^*)\right) + E_g\left(\log g(\underset{\sim}{y}^*)\right). \qquad (1)$$

Remark. $KL(g, g) = 0$ and $KL(f_{\underset{\sim}{\theta}}, g) \geq 0$.

since $-\log t$ is a convex function, leading to $E(-\log X) \geq -\log E(X)$

by Jensen's inequality.

(2) More specifically, letting $X = \frac{f_{\underset{\sim}{\theta}}(\underset{\sim}{y}^*)}{g(\underset{\sim}{y}^*)}$, we have by

Jensen's inequality and the convexity of $-\log t$,

$$E_g\left(-\log \frac{f_{\underset{\sim}{\theta}}(\underset{\sim}{y}^*)}{g(\underset{\sim}{y}^*)}\right) \geq -\log E_g\left(\frac{f_{\underset{\sim}{\theta}}(\underset{\sim}{y}^*)}{g(\underset{\sim}{y}^*)}\right).$$

g is true

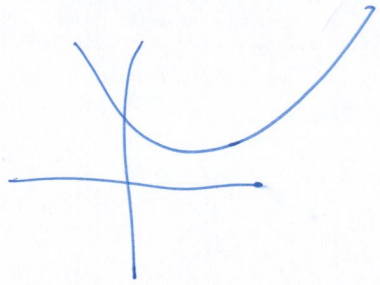
Moreover,

$$E_g\left(\frac{f_{\underset{\sim}{\theta}}(\underset{\sim}{y}^*)}{g(\underset{\sim}{y}^*)}\right) = \int \frac{f_{\underset{\sim}{\theta}}(\underset{\sim}{y}^*)}{g(\underset{\sim}{y}^*)}\, g(\underset{\sim}{y}^*)\, d\underset{\sim}{y}^*$$

$$= \int f_{\underset{\sim}{\theta}}(\underset{\sim}{y}^*)\, d\underset{\sim}{y}^* = 1$$

because $f_{\underset{\sim}{\theta}}(\underset{\sim}{y}^*)$ is a pdf and hence integrates to "1".

Note that Jensen's inequality says that for any convex function $h(\cdot)$, and random variable X,



$$E(h(X)) \geq h(E(X)),$$

whenever the expectations exist. For example,

$$EX^2 \geq (EX)^2,$$

$$-E(\log X) \geq -\log EX,$$

$$E(e^X) \geq e^{E(X)}$$

$$\vdots$$

⊙ In the derivation of AIC, $\underset{\sim}{y}^*$ is regarded as the "future", whereas $\underset{\sim}{y}$, an independent copy of $\underset{\sim}{y}^*$, denotes the observations at hand.

(namely $\underset{\sim}{y}$ and $\underset{\sim}{y}^*$ are independent, but have the same distribution.)

⊙ Since the factor $E_g(\log g(\underset{\sim}{y}^*))$ in (1) will be cancelled out when comparing across different approximation models (families), the goal here is to construct an asymptotically unbiased estimate of

$$-E_g(\log f_{\hat{\underset{\sim}{\theta}}(\underset{\sim}{y})}(\underset{\sim}{y}^*)),$$

(true model; future; "present")

in which $\hat{\underset{\sim}{\theta}}(\underset{\sim}{y})$ is the MLE of $\underset{\sim}{\theta}$ using observations $\underset{\sim}{y}$.

(5)

⊙ A natural estimate of

$$-E_{\underset{\sim}{g}}\left(\log f_{\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y})}(\underset{\sim}{Y}^*) \right)$$ is given

(replaced by $\underset{\sim}{y}$)

by $$-\log f_{\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y})}(\underset{\sim}{y})$$ ← the log likelihood function of $\underset{\sim}{y}$ based on $f_{\underset{\sim}{\theta}}(\cdot)$.

(unknown; dropped here)

However, a "bias correction" term is needed to achieve asymptotic unbiasedness.

(6)

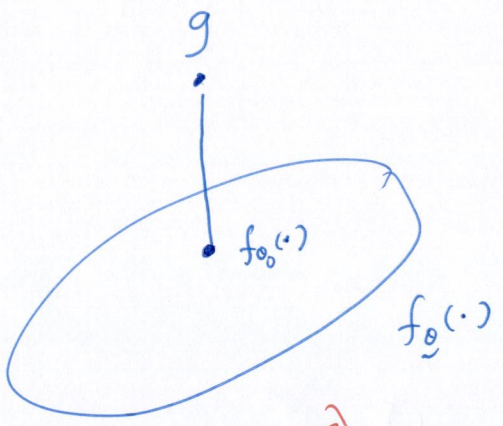
① To find the bias correction term, note first that by Taylor's theorem,

$$E_g\left(\log f_{\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y})}(\underset{\sim}{Y}^*)\right)$$

$$= E_g\left(\log f_{\underset{\sim}{\theta}_0}(\underset{\sim}{Y}^*)\right)$$

$$+ E_g\left[\left(\frac{\partial}{\partial \underset{\sim}{\theta}}\log f_{\underset{\sim}{\theta}_0}(\underset{\sim}{Y}^*)\right)'\left(\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y})-\underset{\sim}{\theta}_0\right)\right]$$

$$+ \frac{1}{2}\, E_g\left[\left(\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y})-\underset{\sim}{\theta}_0\right)\left(\frac{\partial^2 \log f_{\underset{\sim}{\theta}^*}(\underset{\sim}{Y}^*)}{\partial\theta_i\,\partial\theta_j}\right)\left(\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y})-\underset{\sim}{\theta}_0\right)\right], \qquad (2)$$



where $\underset{\sim}{\theta}_0 = \arg\min_{\underset{\sim}{\theta}\in\Theta} KL(f_{\underset{\sim}{\theta}}, g)$ and $\|\underset{\sim}{\theta}^* - \underset{\sim}{\theta}_0\| \le \|\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y}) - \underset{\sim}{\theta}_0\|$.

(7)

⊙ Assume $f_{\theta_0} = g$ (the true model is included among the approximation family.)

(2) becomes

(3)

$$E_g\left(\log f_{\hat{\theta}(Y)}(Y^*)\right)$$

$$= E_{f_{\theta_0}}\left(\log f_{\theta_0}(Y^*)\right) + E_{f_{\theta_0}}\left[\left(\frac{\partial}{\partial \theta}\log f_{\theta_0}(Y^*)\right)'\left(\hat{\theta}(Y)-\theta_0\right)\right]$$

$$+ \frac{1}{2} E_{f_{\theta_0}}\left[\left(\hat{\theta}(Y)-\theta_0\right)' \frac{\partial^2 \log f_{\theta^*}(Y^*)}{\partial\theta_i\,\partial\theta_j}\left(\hat{\theta}(Y)-\theta_0\right)\right]$$

$$\approx E_{f_{\theta_0}}\left(\log f_{\theta_0}(Y^*)\right) + \frac{1}{2} E_{f_{\theta_0}}\left[\left(\hat{\theta}(Y)-\theta_0\right)' \frac{\partial^2 \log f_{\theta_0}(Y^*)}{\partial\theta_i\,\partial\theta_j}\left(\hat{\theta}(Y)-\theta_0\right)\right]$$

because $\theta^* \approx \theta_0$ and $E_{f_{\theta_0}}\left[\left(\frac{\partial}{\partial\theta}\log f_{\theta_0}(Y^*)\right)'\left(\hat{\theta}(Y)-\theta_0\right)\right]$

Y^* and Y are independent $= E_{f_{\theta_0}}\left(\frac{\partial}{\partial\theta}\log f_{\theta_0}(Y^*)\right) E\left[\left(\hat{\theta}(Y)-\theta_0\right)\right]$

$E_{f_{\theta_0}}\left(\frac{\partial}{\partial\theta}\log f_{\theta_0}(Y^*)\right) = 0$

(8) ⊙ Moreover, assume $\lim_{n\to\infty} \frac{1}{n} E_{f_{\underset{\sim}{\theta}_0}}\left(-\frac{\partial^2 \log f_{\underset{\sim}{\theta}_0}(\underset{\sim}{Y}^*)}{\partial\theta_i \partial\theta_j}\right)$ ($\underset{\sim}{Y}^*$ can also be replaced by $\underset{\sim}{Y}$)

$= I(\underset{\sim}{\theta}_0)$ exist.

Then, we have shown (heuristically) that

$$\sqrt{n}\left(\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y}) - \underset{\sim}{\theta}_0\right) \xrightarrow{d} N\left(\underset{\sim}{0}, I^{-1}(\underset{\sim}{\theta}_0)\right),$$

and

$$-\frac{1}{n}\frac{\partial^2 \log f_{\underset{\sim}{\theta}_0}(\underset{\sim}{Y}^*)}{\partial\theta_i \partial\theta_j} \xrightarrow{pr.} +I(\underset{\sim}{\theta}_0),$$

leading to

$$-\left(\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y}) - \underset{\sim}{\theta}_0\right)' \frac{\partial^2 \log f_{\underset{\sim}{\theta}_0}(\underset{\sim}{Y}^*)}{\partial\theta_i \partial\theta_j} \left(\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y}) - \underset{\sim}{\theta}_0\right) \xrightarrow{d} \chi^2(p),$$

assuming $\underset{\sim}{\theta}$ is a p-dimensional vector.

(9)

As a result, (3) becomes

$$E_{f_{\theta_0}}\left(\log f_{\hat{\theta}(\underset{\sim}{Y})}(\underset{\sim}{Y}^*)\right)$$

$$\simeq E_{f_{\theta_0}}\left(\log f_{\theta_0}(\underset{\sim}{Y}^*)\right) - p/2$$

this replacement won't change anything

$$= E_{f_{\theta_0}}\left(\log f_{\theta_0}(\underset{\sim}{Y})\right) - p/2 \qquad (X)$$

(10)

(a)

By Taylor's expansion again,

$$\log f_{\underset{\sim}{\theta}_0}(\underset{\sim}{Y}) = \log f_{\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y})}(\underset{\sim}{Y}) + \left(\frac{\partial}{\partial \underset{\sim}{\theta}} \log f_{\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y})}(\underset{\sim}{Y})\right)\left(\underset{\sim}{\theta}_0 - \hat{\underset{\sim}{\theta}}(\underset{\sim}{Y})\right)$$

$$+ \frac{1}{2}\left(\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y}) - \underset{\sim}{\theta}_0\right)\left(\frac{\partial^2 \log f_{\underset{\sim}{\theta}}(\underset{\sim}{Y})}{\partial\theta_i \partial\theta_j}\right)\left(\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y}) - \underset{\sim}{\theta}_0\right),$$

where $\| \underset{\sim}{\theta} - \underset{\sim}{\theta}_0 \| \leq \| \hat{\underset{\sim}{\theta}}(\underset{\sim}{Y}) - \underset{\sim}{\theta}_0 \|$. since

$\frac{\partial}{\partial \underset{\sim}{\theta}} \log f_{\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y})}(\underset{\sim}{Y}) = \underset{\sim}{0}$ (why?) and

$$-\left(\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y}) - \underset{\sim}{\theta}_0\right) \frac{\partial^2 \log f_{\underset{\sim}{\theta}}(\underset{\sim}{Y})}{\partial\theta_i \partial\theta_j} \left(\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y}) - \underset{\sim}{\theta}_0\right) \xrightarrow{\mathcal{L}} -\left(\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y}) - \underset{\sim}{\theta}_0\right) \frac{\partial^2 \log f_{\underset{\sim}{\theta}_0}(\underset{\sim}{Y})}{\partial\theta_i \partial\theta_j} \left(\hat{\underset{\sim}{\theta}}(\underset{\sim}{Y}) - \underset{\sim}{\theta}_0\right)$$

$\xrightarrow{\mathcal{L}} \chi^2(p)$, it follows that

(11)

$$E_{f_{\theta_0}}\left(\log f_{\theta_0}(Y)\right) \doteq E_{f_{\theta_0}}\left(\log f_{\hat{\theta}(Y)}(Y)\right) - p/2 \qquad (5)$$

(*) Combining (X) and (5) yields

$$\text{[illegible]} - E_{f_{\theta_0}}\left(\log f_{\hat{\theta}(Y)}(Y^*)\right)$$

$$\rightarrow \doteq E_{f_{\theta_0}}\left(-\log f_{\hat{\theta}(Y)}(Y) + p\right) \qquad (6)$$

This is what we call

~~asymptotically~~ "asymptotic unbiasedness"

This is our bias correction term!!

(12) Now, the definition of AIC is given by

$$-2 \log f_{\hat{\underset{\sim}{\theta}}(\underset{\sim}{x})} + 2p \,,$$

which is the quantity inside the expectation on the RHS of (6) multiplied by "2".

a8.1

• 选择模型的评价：

— 致性（Consistency）：

真实模型被包括在候选模型当中

时的计算量.

渐近有效性（Asymptotic efficiency）.

当候选模型不包含真模时

的计算量.

(a9)

- Consistency

the smallest true model

$$P(\hat{J} = J_0) \longrightarrow 1 \quad \text{as } n \to \infty$$

- Asymptotic efficiency:

$$E\left(y_{n+1} - \hat{\beta}_{\hat{J}}' \underset{\sim}{x}_i(\hat{J})\right)^2$$

achieves the minimum value among

$$E\left(y_{n+1} - \hat{\beta}_{J}' \underset{\sim}{x}_i(J)\right)^2$$

asymptotically

(A10)

Consistent criteria:

BIC, HQ, Delete-n_v-out CV, $n_v \to \infty$,

APE

Asymptotically efficient criteria.

特殊限制：候選變數之個數隨樣本數增加而遞增至 ∞.

AIC, C_p, APE_{δ_n}, $\delta_n \to 1$,

Mallows' C_p, noting that

APE_{δ_n} is asymptotically equivalent to

the information criterion IC_{p_n} with $p_n = 1 + \dfrac{\log \delta_n^{-1}}{1-\delta_n}$.

The theorem shows that the increase in risk corresponding to the use of $\hat{k}$ depends asymptotically only on σ_ξ^2 and $K-k_0$. From the proof we can easily obtain the bounds

$$0 \leqslant \lim_{N\to\infty} NE[L\{a,\hat{a}(\tilde{k})\}-L\{a,\hat{a}(k_0)\}] \leqslant (K-k_0)\sigma_\xi^2,$$

where $\tilde{k}$ is any selection method such that

$$E\{\|\hat{a}(\tilde{k})-\hat{a}(k_0)\|^2 I_{(\tilde{k}\leqslant k_0)}\} = o(1/N).$$

The lower bound is attained if k_0 is known and $\tilde{k}=k_0$, and the upper bound is attained if the trivial selection $\tilde{k}=K$ is taken. The numerical examples in Table 2 show that if $K-k_0$ is small we have not so much gain but if it becomes large we have a relatively large gain compared to the above trivial selection.

Table 2. *The increase in risk by using Akaike's information criterion* ($\sigma_\xi^2=1$)

$K-k_0$	1	2	3	4	5	6	7	8	9	10
Increase	0·5724	0·9784	1·285	1·523	1·711	1·863	1·985	2·084	2·200	2·330
Increase/$(K-k_0)$	0·5724	0·4892	0·4283	0·3808	0·3422	0·3105	0·2836	0·2605	0·2444	0·2330

Table 3. *The frequency of the order selected by Akaike's information criterion in* 500 *realizations of a first-order process*

N	a_1 \ k	0	1	2	3	4	5	6	7	8	9	10
50	−0·8	0	366	61	29	24	5	5	6	2	2	0
	−0·3	100	281	48	36	18	6	4	4	3	0	0
	+0·3	128	254	48	27	24	7	4	3	4	1	0
	+0·8	0	369	58	32	16	9	7	4	1	3	1
100	−0·8	0	365	61	33	20	5	11	1	1	1	2
	−0·3	30	333	65	25	16	12	8	8	2	0	1
	+0·3	30	322	76	33	17	9	6	3	3	1	0
	+0·8	0	355	68	37	16	9	6	5	4	0	0
200	−0·8	0	370	50	26	25	5	10	6	4	1	3
	−0·3	1	361	54	30	14	12	8	8	5	2	5
	+0·3	1	358	58	35	14	12	7	7	5	3	0
	+0·8	0	360	56	35	22	4	9	7	2	1	4
300	−0·8	0	359	61	32	16	8	11	3	6	3	1
	−0·3	0	358	61	35	12	15	6	6	3	2	2
	+0·3	0	361	67	32	13	8	5	7	5	2	0
	+0·8	0	354	65	34	23	7	6	3	5	2	1
400	−0·8	0	362	55	36	20	9	4	4	7	2	1
	−0·3	0	373	61	20	14	9	9	7	3	2	2
	+0·3	0	373	57	25	15	13	5	5	2	2	3
	+0·8	0	356	60	41	16	13	3	2	2	2	5
500	−0·8	0	366	61	35	12	10	7	2	3	3	1
	−0·3	0	372	62	21	15	11	4	7	6	1	1
	+0·3	0	375	59	20	15	16	5	5	2	2	1
	+0·8	0	376	50	32	21	7	6	2	5	1	0
Asymptotic (Table 1)		0	359	57	29	18	12	8	6	4	4	3

for a test of significance for $\hat{\rho}^2(k_0+1|k_0)$ would be about 4. For this reason also the value $c = 1$ was used since it would seem pedantic, for the values of N used in Table 1, to choose some value of c such as 1·01. In Table 1 below the results of 100 replications of a number of simulations, for varying N and varying $\alpha(1)$ (k_0 being unity), are presented.

TABLE 1

Frequencies of estimated order. First-order autoregression

		N									
		50		100		200		500		1000	
α	k	ϕ	*AIC*	ϕ	*AIC*	ϕ	*AIC*	ϕ	*AIC*	ϕ	*AIC*
0·1	0	85	72	79	59	55	41	30	17	9	2
	1	9	10	16	22	40	37	66	57	86	60
	2	2	5	5	10	2	9	3	14	4	17
	3	1	4	0	2	2	4	1	7	1	4
	>3	3	9	0	7	1	9	0	5	0	17
0·3	0	35	24	8	5	2	1	0	0	0	0
	1	56	51	87	72	95	79	90	70	94	74
	2	5	9	4	11	2	7	7	13	5	11
	3	1	6	1	4	0	7	3	7	1	7
	>3	3	10	0	8	1	6	0	10	0	8
0·5	0	1	1	0	0	0	0	0	0	0	0
	1	78	66	90	74	91	74	93	69	93	70
	2	14	17	7	11	6	14	5	11	6	12
	3	5	8	1	4	1	8	2	4	1	4
	>3	2	8	2	11	2	4	0	16	0	14
0·7	0	1	0	0	0	0	0	0	0	0	0
	1	78	67	91	72	93	78	94	78	95	70
	2	19	21	5	10	3	9	6	11	5	17
	3	1	6	2	7	3	5	0	5	0	2
	>3	1	6	2	11	1	8	0	6	0	11

The table shows what is to be expected, namely underestimation of the order relative to AIC for smaller N and smaller α but better results than for AIC for larger N or larger α.

In a second simulation we took $x(n) = \varepsilon(n)+\varepsilon(n-1)$ and again used $\phi(k)$ and AIC(k). The results of 100 replications for $N = 100$ are shown in Table 2.

TABLE 2

Frequencies of estimated order. First-order moving average

	k											
	0	1	2	3	4	5	6	7	8	9	10	>10
ϕ	0	0	10	19	30	12	15	5	7	1	1	0
AIC	0	0	2	3	19	9	22	13	15	4	7	6

Here $\rho(k|k-1) = (-1)^k/k$. Since $\{2N^{-1}\ln\ln N\}^{\frac{1}{2}} = 0{\cdot}18$, $(2N^{-1})^{\frac{1}{2}} = 0{\cdot}14$ it is evident that the results are much as might be expected.

TABLE 1
Empirical estimates of $PE(\hat{x}_{n+1}(\hat{k}_n^A))$ *and* $\gamma_{\text{opt}}(n, K_n)$

		θ_0							
		0.8		**0.6**		**−0.6**		**−0.8**	
ϕ_0	n/K_n	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
−0.9	60/7	1.23	1	1.46	1	1.64	1	1.68	1
	120/10	1.21	0.56	1.46	0.54	1.58	0.58	1.55	0.71
	200/14	1.29	0.36	1.56	0.34	1.67	0.38	1.52	0.54
	500/22	1.25	0.17	1.56	0.15	1.63	0.17	1.53	0.34
	1000/31	1.29	0.09	1.54	0.08	1.58	0.10	1.46	0.17
−0.7	60/7	1.23	1	1.44	1	2.90	1	2.49	1
	120/10	1.25	0.57	1.49	0.54	2.48	0.68	2.37	0.70
	200/14	1.29	0.37	1.61	0.35	2.25	0.51	1.95	0.59
	500/22	1.31	0.17	1.53	0.16	1.97	0.27	1.65	0.38
	1000/31	1.28	0.10	1.56	0.09	1.93	0.17	1.62	0.23
−0.5	60/7	1.23	1	1.50	1	3.48	1	1.49	1
	120/10	1.25	0.57	1.62	0.53	3.01	0.65	1.47	0.67
	200/14	1.33	0.36	1.57	0.35	2.78	0.47	1.53	0.46
	500/22	1.33	0.17	1.56	0.16	2.29	0.27	1.44	0.21
	1000/31	1.26	0.10	1.56	0.09	1.99	0.17	1.42	0.13
0.5	60/7	1.55	1	3.10	1	1.49	1	1.25	1
	120/10	1.51	0.67	2.98	0.63	1.59	0.55	1.23	0.56
	200/14	1.48	0.46	2.86	0.45	1.55	0.38	1.31	0.37
	500/22	1.47	0.22	2.45	0.26	1.61	0.16	1.28	0.17
	1000/31	1.41	0.13	1.99	0.16	1.57	0.09	1.32	0.10
0.7	60/7	2.71	1	2.97	1	1.55	1	1.25	1
	120/10	2.31	0.71	2.56	0.62	1.58	0.53	1.25	0.56
	200/14	1.92	0.62	2.31	0.48	1.61	0.36	1.29	0.37
	500/22	1.79	0.37	2.04	0.27	1.53	0.16	1.28	0.18
	1000/31	1.56	0.24	1.95	0.16	1.44	0.09	1.31	0.10
0.9	60/7	1.75	1	1.58	1	1.43	1	1.24	1
	120/10	1.56	0.70	1.61	0.57	1.50	0.53	1.23	0.57
	200/14	1.57	0.51	1.66	0.37	1.58	0.32	1.31	0.37
	500/22	1.49	0.29	1.68	0.17	1.54	0.15	1.31	0.17
	1000/31	1.48	0.17	1.57	0.10	1.47	0.08	1.29	0.10

NOTE. Column P denotes the empirical estimates of $PE(\hat{x}_{n+1}(\hat{k}_n^A))$ and column R denotes the empirical estimates of $\gamma_{\text{opt}}(n, K_n)$.

than those in the other cases in this category. However, the reduction in the values of $\widehat{PE}(\hat{x}_{n+1}(\hat{k}_n^A))$ is also much smaller (only a slightly decreasing trend can be observed). Another observation regarding this category is that, as (n, K_n) increases to (1000, 31), $\widehat{PE}(\hat{x}_{n+1}(\hat{k}_n^A))$ decreases to a value around 1.5 if $\theta_0 = \pm 0.8$, and decreases to a value around 1.95 if $\theta_0 = \pm 0.6$. The third category contains the remaining parameter combinations, namely, $(\phi_0, \theta_0) = (0.5, 0.8)$ and $(-0.5, -0.8)$.

all

AIC 與 BIC 的戰爭：

起因：我們永遠不會知道

真模型是否被含括在

候選模型裡，故無

法判斷要用 AIC 或 BIC.

解決之道：AIC 與 BIC 大和解

model selection consistency and minimax-rate optimality in estimating the regression function cannot be resolved. But this does not indicate that there exists no criterion achieving the pointwise asymptotic efficiency in both well-specified and mis-specified scenarios, because the minimaxity (uniformity over the linear coefficients) is intrinsically different from the (pointwise) efficiency. In the remarkable work by Ing (2007), a hybrid selection procedure combining AIC and a BIC-like criterion was proposed. Loosely speaking, if a BIC-like criterion selects the same model at sample sizes N^ℓ $(0 < \ell < 1)$ and N, then with high probability (for large N) the model class is well-specified and the true model has been converged to, and thus a BIC-like criterion is used; otherwise AIC is used. Under some conditions, the hybrid criterion was proved to achieve the pointwise asymptotic efficiency in both well-specified and mis-specified scenarios. In estimating regression functions with independent observations, Yang (2007) proposed a similar approach to adaptively achieve asymptotic efficiency for both parametric and nonparametric situations, by examining whether BIC selects the same model again and again at different sample sizes (instead of only two sample sizes used by Ing (2007)). Liu & Yang (2011) proposed a method to adaptively choose between AIC and BIC based on a measure called parametricness index. In the context of sequential Bayesian model averaging, Erven, Grünwald & De Rooij (2012) and van der Pas & Grünwald (2014) used a switching distribution to encourage early switch to a better model and offered interesting theoretical understanding on its simultaneous properties. Cross-validation has also been proposed as a general solution to choosing between AIC and BIC. It was shown by Zhang & Yang (2015) that, with a suitably chosen data splitting ratio, the composite criterion asymptotically behaves like the better one of AIC and BIC for both the AIC and BIC territories.

In this paper, we introduce a new model selection criterion which is referred to as the bridge criterion (BC) for autoregressive models. The bridge criterion is able to address the following two issues: First, given a realistic time series data, an analyst is usually unaware of whether the model class is well-specified or not; Second, even if the model class is known

TABLE 1
Empirical estimates of $RE(\hat{k}_n)$

	Models	APE_{δ_n}				IC		Two-stage		
n	(ϕ_0, θ_0)	$\delta_{1,n}$	$\delta_{2,n}$	$\delta_{3,n}$	$\delta_{4,n}$	HQ	BIC	$n^{0.69}$	$n^{0.72}$	$n^{0.75}$
180	(0.0, 0.98)	0.88	0.93	0.92	0.93	0.89	0.78	0.95	0.94	0.94
	(0.5, 0.8)	0.95	0.95	0.95	0.94	0.98	0.83	0.98	0.97	0.97
	(0.5, 0.4)	1.28	1.07	1.05	1.03	1.36	1.26	1.08	1.08	1.08
	(0.9, 0.0)	2.21	1.34	1.33	1.28	2.31	3.59	1.81	1.86	1.95
300	(0.0, 0.98)	0.88	0.94	0.94	0.94	0.89	0.74	0.97	0.96	0.95
	(0.5, 0.8)	0.98	0.99	0.98	0.97	0.96	0.79	0.95	0.94	0.93
	(0.5, 0.4)	1.28	1.03	1.03	1.03	1.24	1.24	1.09	1.09	1.09
	(0.9, 0.0)	2.18	1.37	1.32	1.26	2.44	3.46	1.95	1.99	2.07
500	(0.0, 0.98)	0.85	0.94	0.95	0.95	0.85	0.68	0.96	0.95	0.94
	(0.5, 0.8)	0.97	0.97	0.97	0.96	0.98	0.78	0.97	0.95	0.95
	(0.5, 0.4)	1.28	1.10	1.05	1.04	1.32	1.17	1.03	1.02	1.06
	(0.9, 0.0)	2.31	1.36	1.31	1.27	2.64	4.17	2.39	2.43	2.41
1000	(0.0, 0.98)	0.86	0.95	0.96	0.95	0.86	0.66	0.99	0.98	0.98
	(0.5, 0.8)	1.05	0.97	0.96	0.96	1.01	0.80	0.97	0.97	0.95
	(0.5, 0.4)	1.36	1.12	1.09	1.04	1.37	1.08	1.00	1.00	0.98
	(0.9, 0.0)	2.33	1.27	1.26	1.21	2.86	4.07	2.65	2.74	2.67

Note: $\delta_{1,n} = (\log n)^{-1}$, $\delta_{2,n} = 1 - (2/3)(\log n)^{-0.1}$, $\delta_{3,n} = 1 - (2/3)(\log n)^{-0.12}$, and $\delta_{4,n} = 1 - (2/3)(\log n)^{-0.14}$.

(2) HQ and $APE_{\delta_{1,n}}$, where $\delta_{1,n} = (\log n)^{-1}$. First note that the prediction efficiencies of these two criteria seem quite close. They perform comparably to AIC when $\theta_0 = 0.8$, and much better than it when $\theta_0 \leq 0.4$. This phenomenon can be explained by the fact that HQ and $APE_{\delta_{1,n}}$ are asymptotically efficient in both the finite-order AR model and the AR(∞) model with AR coefficients decaying exponentially (see Theorem 5). Their efficiencies, however, are smaller than AIC in the case $\theta_0 = 0.98$. Since it is difficult to distinguish between an MA(1) process with a very large MA coefficient and an AR(∞) process with AR coefficients decaying algebraically in finite samples, Examples 3 and 8 (which show that HQ and $APE_{\delta_{1,n}}$ are not asymptotically efficient in the algebraic-decay case) may explain why HQ and $APE_{\delta_{1,n}}$ perform worse than AIC when θ_0 is very close to unity. In addition, we also observe that these two criteria are not as efficient as BIC in the case $\theta_0 = 0$, but they beat BIC in all other cases.

(3) $APE_{\delta_{i,n}}$, $i = 2, 3, 4$, where $\delta_{2,n} = 1 - (2/3)(\log n)^{-0.1}$, $\delta_{3,n} = 1 - (2/3) \times (\log n)^{-0.12}$ and $\delta_{4,n} = 1 - (2/3)(\log n)^{-0.14}$. Table 1 shows that $APE_{\delta_{i,n}}$, $i = 2, 3, 4$, holds a slight advantage (disadvantage) over AIC when $\theta_0 = 0.4$ ($\theta_0 \geq 0.8$). However, since the amount of the advantage (disadvantage) is not sizable, these Monte Carlo results seem to support the theoretical findings revealed in Examples 4–6 and 8 that AIC and these APE_{δ_n} criteria are asymptotically equivalent in

ⓐ12

另一場方興未艾的戰爭

源起：第一，我們不被容許隨著樣本數之增加而增加估計參數的個數，~~[illegible]~~

~~[illegible]~~現有的漸近有效準則皆失去漸近有效性.

913

解决之道：MRIC

Misspecification-resistant information criterion

抗误识息准则

Inspired by (3), our strategy to achieve (22) is to first construct the method of moments estimators of $\mathrm{MI}_h(l)$ and $L_h(l)$,

$$\hat{\sigma}_h^2(l) = N^{-1}\sum_{t=1}^{N}\left(y_{t+h} - \hat{\boldsymbol{\beta}}_{n,l}^{\top}(h)\mathbf{x}_t(J_l)\right)^2 \equiv N^{-1}\sum_{t=1}^{N}(\hat{\varepsilon}_{t,h}^{(l)})^2,$$

and

$$\widehat{L}_h(l) = \mathrm{tr}\left(\widehat{\mathbf{R}}_N^{-1}(l)\widehat{\mathbf{C}}_{h,0}(l)\right) + 2\,\mathrm{tr}\left(\sum_{s=1}^{h-1}\widehat{\mathbf{R}}_N^{-1}(l)\widehat{\mathbf{C}}_{h,s}(l)\right),$$

respectively, where $\widehat{\mathbf{R}}_N(l) = N^{-1}\sum_{t=1}^{N}\mathbf{x}_t(J_l)\mathbf{x}_t^{\top}(J_l)$ and

$$\widehat{\mathbf{C}}_{h,s}(l) = (N-s)^{-1}\sum_{t=1}^{N-s}\mathbf{x}_t(J_l)\mathbf{x}_{t+s}^{\top}(J_l)\hat{\varepsilon}_{t,h}^{(l)}\hat{\varepsilon}_{t+s,h}^{(l)}.$$

We then use h-step MRIC, $\mathrm{MRIC}_h(l)$, to quantify the performance of J_l, where

$$\mathrm{MRIC}_h(l) = \hat{\sigma}_h^2(l) + \frac{C_n}{n}\widehat{L}_h(l), \tag{23}$$

with

$$\frac{C_n}{n^{1/2}} \to \infty, \tag{24}$$

and

$$\frac{C_n}{n} \to 0. \tag{25}$$

Finally, we choose model $J_{\hat{l}_h}$, in which $\hat{l}_h = \arg\min_{1\le l\le K}\mathrm{MRIC}_h(l)$. The major difference between $\mathrm{MRIC}_h(l)$ and the natural estimator $\hat{\sigma}_h^2(l) + n^{-1}\widehat{L}_h(l)$ of $\mathrm{E}\,(y_{n+h} - \hat{y}_{n+h}(l))^2$ (cf.(3)) is that the former contains an additional penalty factor C_n. This factor plays a crucial role in search of the best predictive model and is particularly relevant in situations where several competing models share the same MI. To see this, note first that under (C1)–(C6) and two additional assumptions, (31) and (32) (see below), we have

$$\hat{\sigma}_h^2(l) = \mathrm{MI}_h(l) + O_p(n^{-1/2}), \tag{26}$$

and

$$\widehat{L}_h(l) = L_h(l) + o_p(1), \tag{27}$$

yielding

$$\mathrm{MRIC}_h(l) = \mathrm{MI}_h(l) + O_p(n^{-1/2}) + \frac{C_n}{n}L_h(l) + o_p\left(\frac{C_n}{n}\right). \tag{28}$$

In view of (25), property (28) immediately implies

$$\lim_{n\to\infty} P(J_{\hat{l}_h} \in M_1) = 1.$$

Moreover, it follows from (28) and (24) that for $J_{l_1}, J_{l_2} \in M_1$ with $L_h(l_1) \neq L_h(l_2)$,

$$\lim_{n\to\infty} P(\mathrm{sign}(\mathrm{MRIC}_h(l_1) - \mathrm{MRIC}_h(l_2)) = \mathrm{sign}(L_h(l_1) - L_h(l_2))) = 1, \tag{29}$$

5. Numerical studies

In this section, the performance of MRIC is illustrated via three simulated examples. The first and second examples focus on linear and nonliner models, respectively, whereas the third one addresses high-dimensional models. Throughout this section, the C_n in MRIC is set to n^{α_m} for some $\alpha_m > 0.5$.

Example 1. Let the data be generated according to the following true DGPs.

$$y_{t+1} = \beta_1 z_t + \beta_2 w_t + \varepsilon_{t+1}, \tag{53}$$

in which $\varepsilon_t \sim \text{NID}(0,1)$, $z_t = \phi z_{t-1} + \eta_t$ is a stationary AR(1) process, and $w_t = \theta_1 w_{t-1} + \theta_2 w_{t-2} + \delta_t$ is a stationary AR(2) process, with $\eta_t \sim \text{NID}(0, \sigma_\eta^2)$, $\delta_t \sim \text{NID}(0, \sigma_w^2)$, and $\{\eta_t\}$, $\{\delta_t\}$ and $\{\varepsilon_t\}$ mutually independent. We also let

$$\sigma_\eta^2 = 1 - \phi^2, \sigma_w^2 = 1 - \theta_2^2 - \{\theta_1^2(1+\theta_2)/(1-\theta_2)\},$$

$\beta_1 = \beta_2 = 1$, and $\phi = \theta_1/(1-\theta_2)$, noting that (53) leads to $\gamma_z(0) = 1 = \gamma_w(0)$, where $\gamma_z(j) = \text{E}(z_t z_{t+j})$ and $\gamma_w(j) = \text{E}(w_t w_{t+j})$. In this study, we consider four different (θ_1, θ_2)'s: (0.15, 0.5), (-0.10, 0.65), (-0.40, -0.60), (0.10, -0.95), which are denoted by DGPs I-IV. With observations up to time n, we are interested in performing h-step prediction, with $h = 2$ and 3, using two candidate models,

$$\begin{aligned} J_1 &: \quad y_{n+h} = \alpha z_n + \varepsilon_{n,h}^{(1)}, \\ J_2 &: \quad y_{n+h} = \beta w_n + \varepsilon_{n,h}^{(2)}, \end{aligned}$$

which are misspecified. The MI and VI of candidate J_l are denoted by $\text{MI}_h(l)$ and $L_h(l)$ with $l = 1, 2$. It is shown in Table 3 that $\text{MI}_2(1) = \text{MI}_2(2)$ in all four DGPs, but $L_2(1) < L_2(2)$ in DGPs I and II and $L_2(1) > L_2(2)$ in DGPs III and IV. Therefore, for the two-step prediction, the better predictive model is J_1 (J_2) under DGP I or II (III or IV). On the other hand, Table 3 reveals that $\text{MI}_3(1) > \text{MI}_3(2)$ in all DGPs, yielding that the better predictive model is always J_2 when $h = 3$. The percentage of MRIC (with $\alpha_m = 0.6$) choosing the better candidate is obtained by using 1,000 simulations for sample sizes $n = 200, 500, 1000, 2000, 3000$; see Table 4 ($h = 2$) and Table 5 ($h = 3$). For the sake of comparison, the corresponding percentages of AIC, BIC, GAIC (Konishi and Kitagawa, 1996), GBIC (Lv and Liu, 2014) and GBIC_p (Lv and Liu, 2014) are also reported in Tables 4 and 5, where for candidate J_l,

$$\begin{aligned} \text{AIC}(l) &= \log \hat{\sigma}_h^2(l) + \frac{2\sharp(J_l)}{n}, \\ \text{BIC}(l) &= \log \hat{\sigma}_h^2(l) + \frac{\sharp(J_l)\log n}{n}, \\ \text{GAIC}(l) &= \log \hat{\sigma}_h^2(l) + \frac{2\text{tr}(\widehat{H}_h(l))}{n}, \\ \text{GBIC}(l) &= \log \hat{\sigma}_h^2(l) + \frac{\sharp(J_l)\log n}{n} - \frac{\log\det(\widehat{H}_h(l))}{n}, \\ \text{GBIC}_p(l) &= \log \hat{\sigma}_h^2(l) + \frac{\sharp(J_l)\log n}{n} + \frac{\text{tr}(\widehat{H}_h(l))}{n} - \frac{\log\det(\widehat{H}_h(l))}{n}, \end{aligned}$$

with

$$\widehat{H}_h(l) = \hat{\sigma}_h^{-2}(l)\widehat{\mathbf{R}}_N^{-1}(l)\widehat{\mathbf{C}}_{h,0}(l),$$

Table 3. The values of $\mathrm{MI}_h(1) - \mathrm{MI}_h(2)$ and $L_h(1) - L_h(2)$ in Example 1, and the corresponding better predictive models

	DGP			
	I	II	III	IV
		$h = 2$		
$\mathrm{MI}_h(1) - \mathrm{MI}_h(2)$	0.000	0.000	0.000	0.000
$L_h(1) - L_h(2)$	-0.716	-0.966	0.959	1.873
The better predictive model	J_1	J_1	J_2	J_2
		$h = 3$		
$\mathrm{MI}_h(1) - \mathrm{MI}_h(2)$	0.269	0.428	0.232	0.882
$L_h(1) - L_h(2)$	*	*	*	*
The better predictive model	J_2	J_2	J_2	J_2

*: $L_h(1) - L_h(2)$ can be neglected.

which is a consistent estimator of $\sigma_h^{-2}(l)\mathbf{R}^{-1}(l)\mathbf{C}_{h,0}(l)$. Note first that $\mathrm{MRIC}(l)$ is asymptotically equivalent to

$$\log \hat{\sigma}_h^2(l) + \frac{C_n \hat{\sigma}_h^{-2}(l)\hat{L}_h(l)}{n},$$

which shares a common first term with these five criteria. On the other hand, by featuring a consistent estimator of VI, $\hat{L}_h(l)$, and a suitable penalty term, C_n, the second term of MRIC readily paves the way for a consistent selection of the better predictive model, whether the MIs of candidate models are equal or not. We also mention that this latter property is, in general, not enjoyed by these five criteria because (i) the trace of $\widehat{\mathbf{R}}_N^{-1}(l)\widehat{\mathbf{C}}_{h,0}(l)$ in $\widehat{H}_h(l)$ is a consistent estimator of VI only when $h = 1$ or observations are independent over time, and (ii) the penalty term $\log n$ used in BIC, GBIC and GBIC_p is too weak when misspecified candidates are non-nested (see Sin and White (1996) and Inoue and Kilian (2006) for related discussion). In fact, the criterion values of GAIC (AIC, BIC, GBIC, GBIC_p) for J_1 and J_2 are expected to be close to each other because $\mathrm{MI}_h(1) = \mathrm{MI}_h(2)$, $\sharp(J_1) = \sharp(J_2)$ and $\mathrm{tr}(\mathbf{R}^{-1}(l)\mathbf{C}_{h,0}(l)) = \sharp(J_l)\mathrm{MI}_h(l)$ (under normality). As shown in Table 4, these five criteria behave like a fair coin to choose between two alternatives, and can only select the better candidate about 50% of the time. In contrast, MRIC has a much higher chance of identifying the better model in this difficult situation. Its percentage falls between 67% and 100%, and tends to increase with the sample size and the value of $|L_2(1) - L_2(2)|$.

When $h = 3$, the two competing candidates have different MIs, and hence it becomes much easier to identify the better one. As shown in Table 5, all criteria perform satisfactorily for all sample sizes $n \geq 200$. While in DGPs I and III, MRIC seems slightly worse than the other criteria for $n = 200$, the corresponding percentages are still over 93%.

Example 2. In this example, we consider the following DGP,

$$y_{t+2} = \frac{1}{1 - aB} x_t + \frac{1}{1 - bB} z_t + \varepsilon_{t+2}, \tag{54}$$

Table 4. Percentage of times, across 1,000 simulations, that the better predictive model between J_1 and J_2 of Example 1 is chosen in the case of $h = 2$

		DGPs			
n	Criteria	I	II	III	IV
200	AIC/BIC	51.50	54.50	48.50	46.30
	GAIC	51.40	54.30	49.00	46.70
	GBIC	51.60	54.40	48.50	45.40
	GBICp	51.60	54.40	48.40	46.00
	MRIC	66.80	73.20	76.70	95.80
500	AIC/BIC	51.10	50.70	47.60	49.00
	GAIC	50.80	50.50	47.30	50.90
	GBIC	51.10	50.50	47.60	47.30
	GBICp	51.10	50.70	47.60	49.10
	MRIC	69.80	74.20	85.30	99.70
1000	AIC/BIC	48.10	53.60	53.00	49.40
	GAIC	48.00	53.00	52.40	50.00
	GBIC	48.10	53.50	52.80	49.20
	GBICp	48.10	53.50	53.00	49.40
	MRIC	74.90	80.80	88.70	100.00
2000	AIC/BIC	50.10	49.70	50.80	49.60
	GAIC	50.10	49.50	50.90	49.20
	GBIC	50.30	49.70	50.90	49.30
	GBICp	50.10	49.70	50.80	49.60
	MRIC	78.20	83.90	92.20	100.00
3000	AIC/BIC	51.40	51.20	49.00	50.40
	GAIC	51.40	51.10	48.90	50.60
	GBIC	51.30	51.20	49.00	50.70
	GBICp	51.40	51.20	49.00	50.40
	MRIC	79.80	84.90	93.40	100.00

High Dimensional Regressions

Consider

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, i = 1, 2, \cdots, n,$$

where $\log p = o(n)$.

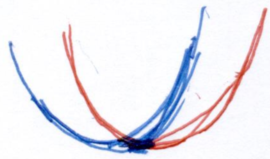
$\star$ Subset selection is infeasible in high-dimensional regression problems.

◇ Two alternatives:

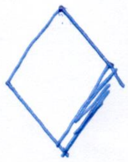
least absolute shrinkage and selection operator

∘ Lasso: Minimizing the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant, that is, $\hat{\boldsymbol{\beta}}^L(\lambda) = (\hat{\beta}_1^L(\lambda), \cdots, \hat{\beta}_p^L(\lambda))^\top$, which is the minimizer of

$$n^{-1}\sum_{t=1}^{n}(y_t - \mathbf{x}_t^\top \mathbf{c})^2 + \lambda\|\mathbf{c}\|_1$$



over $\mathbf{c}$.



||c|| = 1

· [illegible]，一次（滿足三個願望

· 武林至尊，寶刀屠龍，號令天下，莫敢不從

（有神快拜!!）

拼湊出來的倚天劍!

∘ Orthogonal Greedy Algorithm (OGA) (or Pure Greedy Algorithm (PGA)).

Step 1. Define $\mathbf{Y}_n = (y_1, \cdots, y_n)'$ and $\mathbf{x}_j = (x_{1j}, \cdots, x_{nj})'$. Find a variable among $\{\mathbf{x}_1, \cdots, \mathbf{x}_{p_n}\}$ that is most correlated to $\mathbf{Y}_n$. Call the variable $\mathbf{x}_{\hat{j}_1}$ and generate residual $R_1 = \mathbf{Y}_n - M_{\hat{j}_1}\mathbf{Y}_n$, where $M_{\hat{j}_1}$ is the orthogonal projection matrix on span$\{\mathbf{x}_{\hat{j}_1}\}$.

Step 2. Find a variable among $\{\mathbf{x}_1, \cdots, \mathbf{x}_{p_n}\}$ that is most correlated to R_1. Call the variable $\mathbf{x}_{\hat{j}_2}$ and let $R_2 = \mathbf{Y}_n - M_{\hat{j}_1\hat{j}_2}\mathbf{Y}_n$, where $M_{\hat{j}_1\hat{j}_2}$ is the orthogonal projection matrix on span$\{\mathbf{x}_{\hat{j}_1}, \mathbf{x}_{\hat{j}_2}\}$.

⋮

If iterations stop at step m, then the new outcome, $y = y(\mathbf{x}) + \varepsilon$, is predicted by

$$\hat{y}_m(\mathbf{x}) = \hat{\beta}_{\hat{j}_1} x_{\hat{j}_1} + \cdots + \hat{\beta}_{\hat{j}_m} x_{\hat{j}_m},$$

where $\hat{\beta}_{\hat{j}_1}, \cdots, \hat{\beta}_{\hat{j}_m}$ are least squares estimates obtained by regressing y_i on $x_{i\hat{j}_1}, \cdots, x_{i\hat{j}_m}$.

Forward selection

◇ Prediction performance of OGA.

Assumptions:

(C1) $\log p_n = o(n)$.

Moreover, $(\varepsilon_t, \mathbf{x}_t)$ in (1.1) are i.i.d. and such that ε_t is independent of $\mathbf{x}_t$ and

(C2) $E\{\exp(s\varepsilon)\} < \infty$ for $|s| \leq s_0$,

where $(\varepsilon, \mathbf{x})$ denotes an independent replicate of $(\varepsilon_t, \mathbf{x}_t)$. We also assume that $\alpha = 0$ and $E(\mathbf{x}) = \mathbf{0}$. Letting $\sigma_j^2 = E(x_j^2)$, $z_j = x_j/\sigma_j$ and $z_{tj} = x_{tj}/\sigma_j$, we assume that there exists $s > 0$ such that

(C3) $\limsup_{n\to\infty} \max_{1\leq j\leq p_n} E\{\exp(sz_j^2)\} < \infty$.

In addition, we assume the **weak sparsity condition**

(C4) $\sup_{n\geq 1} \sum_{j=1}^{p_n} |\beta_j \sigma_j| < \infty$.

Let

$$\mathbf{\Gamma}(J) = E\{\mathbf{z}(J)\mathbf{z}^\top(J)\},\ \mathbf{g}_i(J) = E(z_i\mathbf{z}(J)),$$

where $\mathbf{z}(J)$ is a subvector of $(z_1, \cdots, z_p)^\top$ and J denotes the associated subset of indices $1, \cdots, p$. We assume that for some positive constant M independent of n,

$$\liminf_{n\to\infty} \min_{1\le\sharp(J)\le K_n} \lambda_{\min}(\mathbf{\Gamma}(J)) > 0, \tag{1}$$

$$\max_{1\le\sharp(J)\le K_n, i\notin J} \|\mathbf{\Gamma}^{-1}(J)\mathbf{g}_i(J)\|_1 < M, \tag{2}$$

where $\sharp(J)$ denotes the cardinality of J and

$$\|\boldsymbol{\nu}\|_1 = \sum_{j=1}^{k} |\nu_j| \text{ for } \boldsymbol{\nu} = (\nu_1, \cdots, \nu_k)^\top.$$

Note that the lhs of (2) is closely related to the "cumulative coherence function" introduced by Tropp (2004).

Theorem 2. *Assume* (C1)-(C4) *and* (1) *and* (2). *Suppose* $K_n \to \infty$ *such that* $K_n = O((n/\log p_n)^{1/2})$. *Then for* OGA,

$$\max_{1 \le m \le K_n} \left(\frac{E[\{y(\mathbf{x}) - \hat{y}_m(\mathbf{x})\}^2 | y_1, \mathbf{x}_1, \cdots, y_n, \mathbf{x}_n]}{m^{-1} + n^{-1} m \log p_n} \right) = O_p(1).$$

◇ A data-driven method of choosing the number of OGA iterations.

Define High-Dimensional Akaike's Information Criterion (HDAIC),

$$\mathrm{HDAIC}(J) = n \log \hat{\sigma}_J^2 + \sharp(J) s \log p_n, \qquad (3)$$

where s is some positive constant, and define

$$\hat{k}_n = \arg \min_{1 \le k \le K_n} \mathrm{HDAIC}(\hat{J}_k),$$

where $\hat{J}_k$ is the index set selected by OGA at the k-th iteration.

We assume that $(\mathbf{x}_t^\top, \varepsilon_t)^\top$ are i.i.d. $(p+1)$-dimensional multivariate normal random vectors with mean $\mathbf{0}$ and covariance matrix

$$\begin{pmatrix} \mathbf{\Gamma} & \mathbf{0} \\ \mathbf{0}^\top & \sigma^2 \end{pmatrix}.$$

Theorem 4 *Under the assumption of Theorem 2 and the normality assumption, it follows that*

$$\frac{E[\{y(\mathbf{x}) - \hat{y}_{\hat{k}_n}(\mathbf{x})\}^2 | y_1, \mathbf{x}_1, \cdots, y_n, \mathbf{x}_n]}{(n^{-1} \log p_n)^{1-(2\gamma)^{-1}}} = O_p(1),$$

provided $s > 14$ in (3).

- **Trim**

Let $\hat{J}_{\hat{k}_n}$ denote the index set chosen by OGA+HDAIC. Since $\hat{J}_{\hat{k}_n}$ may contain irrelevant variables that are included along the OGA path, to exclude irrelevant variables, we make use of HDAIC to define a subset $\hat{N}_n$ of $\hat{J}_{\hat{k}_n}$ by

$$\hat{N}_n = \{\hat{j}_l : \mathrm{HDAIC}(\hat{J}_{\hat{k}_n} - \{\hat{j}_l\}) > \mathrm{HDAIC}(\hat{J}_{\hat{k}_n}), 1 \le l \le \hat{k}_n\} \text{ if } \hat{k}_n > 1,$$

and $\hat{N}_n = \{\hat{j}_1\}$ if $\hat{k}_n = 1$. Note that this refinement only requires the computation of $\hat{k}_n - 1$ additional least squares estimates and their associated residual sum of squares $\sum_{t=1}^{n}(y_t - \hat{y}_{t;\hat{J}_{\hat{k}_n}-\{\hat{j}_l\}})^2, 1 \le l < \hat{k}_n$, in contrast to the intractable combinatorial optimization problem of choosing the subset with the smallest extended BIC among all non-empty subsets of $\{1, \cdots, p_n\}$, for which Chen and Chen (2008, Theorem 1) established variable selection consistency under an "asymptotic identifiability" condition and $p_n = O(n^{\kappa})$ for some $\kappa > 0$.

backward elimination

Table 1. Frequency, in 1,000 simulations, of including all five relevant variables (Correct), of selecting exactly the relevant variables (E), of selecting all relevant variables and i irrelevant variables (E+i), and of selecting the largest model, along the OGA path, which includes all relevant variables (E*).

η	n	p	Method	E	E+1	E+2	E+3	E+4	E+5	E*	Correct	MSPE
0	50	1,000	OGA+HDHQ	812	86	19	8	0	0	1	926	6.150
			OGA+HDBIC	862	52	7	1	0	0	0	922	6.550
			OGA+HDBIC+Trim	919	3	0	0	0	0	0	922	6.550
			OGA+BIC	0	0	0	0	0	0	926	926	8.310
			FR	0	0	0	0	0	0	926	926	8.300
	100	2,000	OGA+HDHQ	993	6	0	1	0	0	0	1,000	0.065
			OGA+HDBIC	999	0	0	1	0	0	0	1,000	0.064
			OGA+HDBIC+Trim	1,000	0	0	0	0	0	0	1,000	0.064
			OGA+BIC	0	0	0	0	0	0	1,000	1,000	1.320
			FR	0	0	0	0	0	0	1,000	1,000	1.729
	200	4,000	OGA+HDHQ	999	1	0	0	0	0	0	1,000	0.034
			OGA+HDBIC	1,000	0	0	0	0	0	0	1,000	0.034
			OGA+HDBIC+Trim	1,000	0	0	0	0	0	0	1,000	0.034
			OGA+BIC	0	0	0	0	0	0	1,000	1,000	0.796
			FR	0	0	0	0	0	0	1,000	1,000	1.612
2	50	1,000	OGA+HDHQ	609	140	36	15	5	2	0	807	13.250
			OGA+HDBIC	629	130	29	5	0	0	0	793	14.110
			OGA+HDBIC+Trim	792	1	0	0	0	0	0	793	14.100
			OGA+BIC	0	0	0	0	0	0	807	807	14.660
			FR	0	0	0	0	0	0	807	807	14.990
	100	2,000	OGA+HDHQ	988	9	3	0	0	0	0	1,000	0.070
			OGA+HDBIC	994	3	3	0	0	0	0	1,000	0.069
			OGA+HDBIC+Trim	1,000	0	0	0	0	0	0	1,000	0.069
			OGA+BIC	0	0	0	0	0	0	1,000	1,000	1.152
			FR	0	0	0	0	0	0	1,000	1,000	1.537
	200	4,000	OGA+HDHQ	1,000	0	0	0	0	0	0	1,000	0.033
			OGA+HDBIC	1,000	0	0	0	0	0	0	1,000	0.033
			OGA+HDBIC+Trim	1,000	0	0	0	0	0	0	1,000	0.033
			OGA+BIC	0	0	0	0	0	0	1,000	1,000	0.779
			FR	0	0	0	0	0	0	1,000	1,000	1.688

straightforward calculations give $\mathbf{c}_{qi} = \eta^2 \mathbf{1}_q$, $\mathbf{R}^{-1}(q) = \mathbf{I} - \{\eta^2/(1+\eta^2 q)\}\mathbf{1}_q \mathbf{1}_q^\top$, and $\operatorname{sign}(\boldsymbol{\beta}(q)) = \mathbf{1}_q$, where $\mathbf{1}_q$ is the q-dimensional vector of 1's. Therefore, for all $i = q+1, \ldots, p$, $|\mathbf{c}_{qi}' \mathbf{R}^{-1}(q) \operatorname{sign}(\boldsymbol{\beta}(q))| = \eta^2 q/(1+\eta^2 q) < 1$. Under (5.6) and some other conditions, Meinshausen and Bühlmann (2006, Thms. 1 and 2) have shown that if $r = r_n$ in the Lasso estimate (3.21) converges to 0 at a rate slower than $n^{-1/2}$, then $\lim_{n\to\infty} P(\hat{L}_n = N_n) = 1$, where $\hat{L}_n$ is the set of regressors whose associated regression coefficients estimated by Lasso(r_n) are nonzero.

Table 2 compares the performance of OGA+HDBIC, OGA+HDHQ and

$(\beta_1, \ldots \beta_5) = (3, -3.5, 4, -2.8, 3.2)$

$\beta_6 = \ldots = \beta_p = 0$ $\quad x_i = d_i + 2w, \; i = 1 \ldots p.$

$\begin{pmatrix} d_1 \\ \vdots \\ d_p \\ w \end{pmatrix} \sim N(0, I).$

Table 2. Frequency, in 1,000 simulations, of including all nine relevant variables and selecting all relevant variables in Example 2; see notation in Table 1.

Method	E	E+1	E+2	E+3	Correct	MSPE
OGA+HDHQ	980	18	1	1	1,000	0.067
OGA+HDBIC	982	16	1	1	1,000	0.067
OGA+HDBIC+Trim	1,000	0	0	0	1,000	0.066
SIS-SCAD	127	19	10	14	289	15.170
ISIS-SCAD	0	0	0	0	1,000	1.486
Adaptive Lasso	1,000	0	0	0	1,000	0.289
LARS	0	0	0	0	1,000	0.549
Lasso	0	0	0	0	1,000	0.625

Table 3. 5-number summaries of squared prediction errors in 1,000 simulations.

Method	Minimum	1st Quartile	Median	3nd Quartile	Maximum
OGA+HDHQ	0.000	0.006	0.026	0.084	1.124
OGA+HDBIC	0.000	0.006	0.026	0.084	1.124
OGA+HDBIC+Trim	0.000	0.005	0.026	0.084	1.124
SIS-SCAD	0.000	0.100	2.498	17.210	507.200
ISIS-SCAD	0.000	0.153	0.664	1.845	21.340
Adaptive Lasso	0.000	0.030	0.118	0.360	3.275
LARS	0.000	0.047	0.224	0.719	5.454
Lasso	0.000	0.067	0.280	0.823	7.399

Using the same notation as in the first paragraph of Example 2, straightforward calculations show that for $q+1 \le j \le p$, $\mathbf{c}_{qj} = (b, \ldots, b)^\top$, $\mathbf{R}(q) = \mathbf{I}$, and $\operatorname{sign}(\boldsymbol{\beta}(q)) = (1, \ldots, 1)^\top$. Therefore, for $q+1 \le j \le p$, $|\mathbf{c}_{qj}^\top \mathbf{R}^{-1}(q)\operatorname{sign}(\boldsymbol{\beta}(q))| = (3q/4)^{1/2} = (7.5)^{1/2} > 1$, and hence (5.6) is violated. On the other hand, it is not difficult to show that (3.2) is satisfied in this example and that

$$\min_{q+1 \le i \le p} |E(x_i y)| > \max_{1 \le i \le q} |E(x_i y)|. \tag{5.8}$$

In fact, $|E(x_i y)| = 24.69$ for all $q+1 \le i \le p$ and $\max_{1 \le i \le q} |E(x_i y)| = \beta_q = 9.75$. Making use of (5.8) and Lemmas A.2 and A.4 in Appendix A, it can be shown that $\lim_{n \to \infty} P(\hat{J}_1 \subseteq \{1, \ldots, q\}) = 0$, and therefore with probability approaching 1, the first iteration of OGA selects an irrelevant variable, which remains in the OGA path until the last iteration.

Table 4 shows that although OGA+HDHQ and OGA+HDBIC fail to choose the smallest set of relevant regressors in all 1,000 simulations, consistent with the above asymptotic theory, they include only 1–3 irrelevant variables while correctly including all relevant variables. Moreover, by using HDBIC to define the subset (4.21) of $\hat{J}_{\hat{k}_n}$, OGA+HDBIC+Trim is able to choose all relevant variables

Table 4. Frequency, in 1,000 simulations, of including all nine relevant variables and selecting all relevant variables in Example 3; see notation in Table 1.

Method	E	E+1	E+2	E+3	Correct	MSPE
OGA+HDHQ	0	39	945	16	1,000	0.035
OGA+HDBIC	0	39	945	16	1,000	0.035
OGA+HDBIC+Trim	1,000	0	0	0	1,000	0.028
SIS-SCAD	0	0	0	0	0	51.370
ISIS-SCAD	0	0	0	0	1,000	0.734
Adaptive Lasso	0	0	0	0	0	27.270
LARS	0	0	0	0	0	0.729
Lasso	0	0	0	0	0	2.283

without including any irrelevant variables in all 1,000 simulations, and its MSPE is close to the oracle value of $q\sigma^2/n = 0.025$ while those of OGA+HDBIC and OGA+HDHQ are somewhat larger. Similar to Example 2, ISIS-SCAD includes all relevant regressors in all 1,000 simulations, but also includes many irrelevant regressors. Its MSPE is 0.73, which is about 29 times the value of $q\sigma^2/n$. The performance of SIS-SCAD is again much worse than that of ISIS-SCAD. It fails to include all relevant regressors in all 1,000 simulations and its MSPE is about 70 times that of ISIS-SCAD.

Like SIS-SCAD, LARS, Lasso, and adaptive Lasso also fail to include all 10 relevant regressors in all 1,000 simulations, even though they also include many irrelevant variables. The smallest number of selected variables in the 1,000 simulations is 8 for adaptive Lasso, 234 for Lasso, and 372 for LARS. The average and the largest numbers of selected variables are 12.59 and 19 for adaptive Lasso, 272.2 and 308 for Lasso, and 393.7 and 399 for LARS. On the other hand, the MSPE of LARS is 0.73, which is about 1/3 of that of Lasso and 1/40 of that of adaptive Lasso. This example shows that when Lasso fails to have the sure screening property, adaptive Lasso, which relies on an initial estimate based on Lasso to determine the weights for a second-stage weighted Lasso, may not be able to improve Lasso and may actually perform worse. The example also illustrates an inherent difficulty with high-dimensional sparse linear regression when irrelevant input variables have substantial correlations with relevant ones. Assumptions on the design matrix are needed to ensure that this difficulty is surmountable; in particular, (3.17) or the second part of (3.2) can be viewed as a "sparsity" constraint, when a candidate irrelevant input variable is regressed on the set of variables already selected by the OGA path, to overcome this difficulty.

6. Concluding Remarks and Discussion

Forward stepwise regression is a popular regression method that seems to be particularly suitable for high-dimensional sparse regression models but has

① 高维统计之现状与未来

- 成功的解决的高维线性回归问题
- 部分解决了高维矩阵估计问题.
- 有许多成功的应用
- 部分涉及非线性问题
- 扩大统计研究领域
- 高维及非参数的问题?
- 高维实时统计?
- 高维序贯方法? [e.g. multi-armed bandit problems with many covariates (contextual bandit problems)]
  - Alpha go (两个棋)
  - digital marketing
  - personalized medicine
- 高维 coordinate descent method ?

⋮

绪论：

关于统计模型选择

特点：
- 实用性高
- 易于产生跨领域合作
- 容易形成研究团队
- 容易吸引非数统背景学生

要求：
- 好的数学背景（逼近理论，分析…）
- 好的私事统计训练（大、小样本理论，参性及非参性模型之预测及推断、贝叶斯理论…）
- 好的计算能力（软件操作、程式，演算法分析…）
- 好的统计直觉？

Overview of semiconductor fabrication

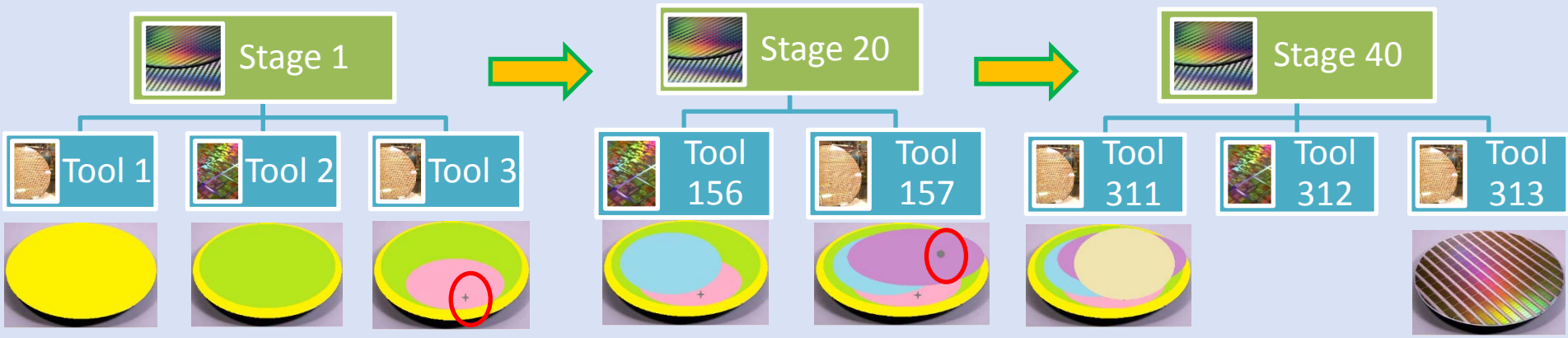
- Multiple stages are involved: ingot growing, sawing into wafers, thermal and local oxidation, photolithography, etching, dopant diffusion, ion implantation, and chemical-mechanical planarization processes.
- Each wafer contain hundreds of dies(or chips), which are separated by scribing and cleaving, then packaged for protection.
- Modern fabs are organized into "workcells" so that all necessary equipment for completing a given stage is placed in the same room, so that chance of mishandling is reduced.

Wafer QA Tests

There are two tests consisted in wafer testing.

- Prior to packaging: Wafer sorting, testing individual dies on the wafer with wafer prober. Failed dies are not packaged, wafers with large proportion of failed dies are discarded.
- After wafer packaging: Final test (functional testing at the completion of production) uses a sample of the packaged wafers to test if the product actually meets the customers' specifications. It provides a concrete application of the methodology developed for identification of defective tools following fault detection.

E-matching 品質管制



- 機台總數量十分龐大，晶圓片數則非常少，特別在製程初期，未大量生產晶圓之時，如何找出製程中，發生問題的機台，此時估計參數大於晶圓片數（即觀察值個數）

  品質控管:影響獲利及競爭力

巨量資料

- **Large p, small n (Big Data)**：p 指的是模型中的解釋變數(機台)之個數，而 n則 指的是樣本數(晶圓片數)。由於變數多樣本少，故在大量解釋變數中挑選出少數真正有影響力的變數有如大海撈針般困難

（西諺：Finding needles in a haystack）

模型

- **Location-dispersion model:**

$$y_i = \beta_0 + \sum_{j=1}^{n_t} \beta_j x_{ij} + \sigma_i \varepsilon_i, \quad \varepsilon_i \quad \text{i.i.d. } N(0,1)$$

$$\sigma_i^2 = \exp\left\{ \alpha_0 + \sum_{j=1}^{n_t} \alpha_j x_{ij} \right\}$$

其中 y_i 是第 i 片晶圓觀測到的電性資料，如WAT值，$i = 1, \cdots, n$，x_{ij} 為第 i 片晶圓通過或不通過第 j 個機台的值，通過為1，不通過為0，β_j 與 α_j 則分別為第 j 個機台在location與dispersion function中的模型係數，$n_t = \sum_{k=1}^{n_s} m_k$ 代表機台總數，n_s 為stage的個數，m_k 為第 k 個stage中的機台個數。

WAT: wafer acceptance

方法

- **三階段選模法：OGA-HDIC-TRIM**

1) OGA: sequentially select the variables
2) HDIC: choose along the OGA path the model that has the smallest value of a suitably chosen criterion $\mathrm{HDIC}(J) = n \log \hat{\sigma}_J^2 + \#(J) w_n \log p$
3) TRIM: exclude irrelevant variables

- 目標：

1) 挑選具影響力機台
2) 部分機台重複使用
3) 決定是否合併資料
4) 提供分析程式工廠上線使用

Case 1 資料說明(2015/09/10)

- 共有245片晶圓
- 每一片所經過的step數與tool數不盡相同。
- 總共step個數為612
- 原始總共tool個數為2740

- 只分析通過『10片晶圓以上、236片晶圓以下』的處理機台
- 剩下的處理機台個數為2504

Case 1 配適模型(2015/09/10)

- 資料配適模型：

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j \in J_b} \hat{\beta}_j x_{ij} + \hat{\sigma}_i \varepsilon_i \quad , \quad \varepsilon_i \text{ are } iid\ N(0,1)$$

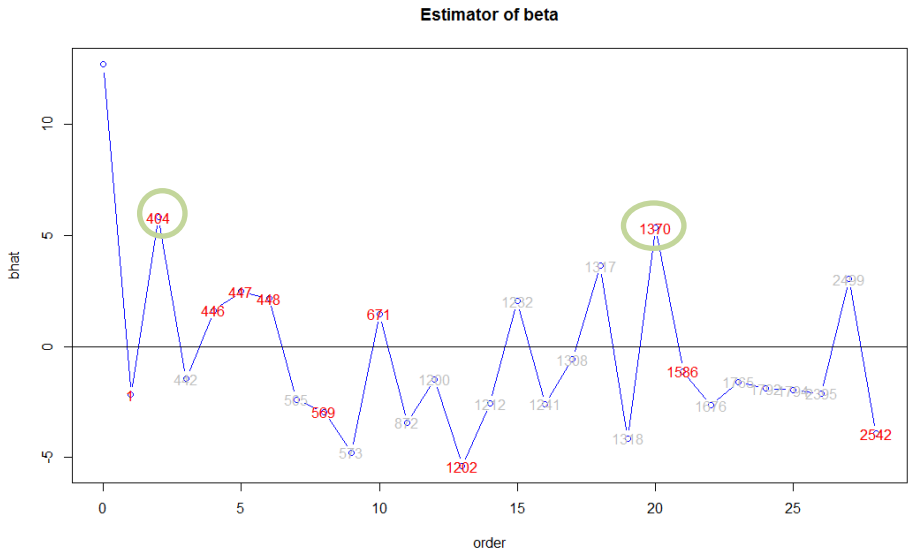
$$\hat{\sigma}_i^2 = \exp\left\{\hat{\alpha}_0 + \sum_{j \in J_a} \hat{\alpha}_j x_{ij}\right\}$$

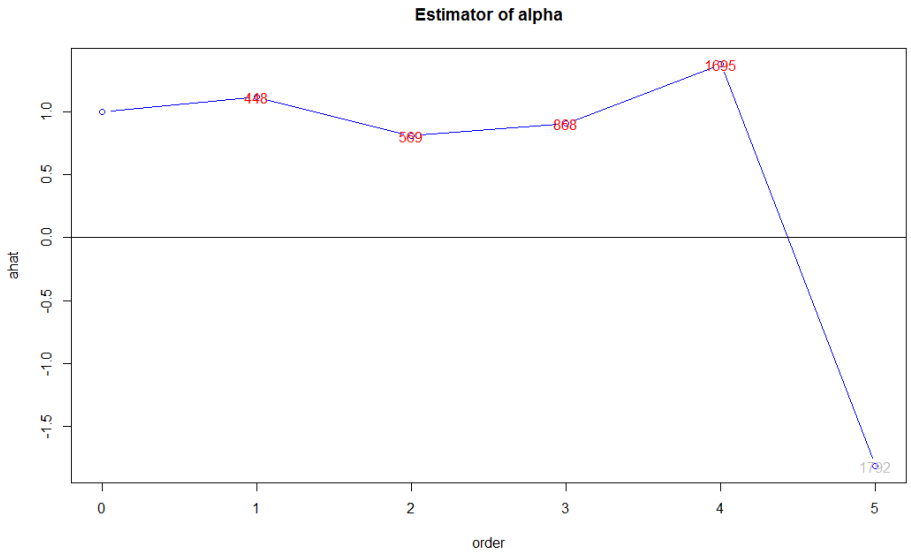
$$J_b = \left\{\begin{array}{c} 404, 1370, 1586, 447, 1202, 569, 2542, 671, 1, 446, \\ 1232, 1765, 1317, 2395, 1676, 1308, 1241, 573, 2499, 448, \\ 505, 1212, 872, 1794, 1318, 442, 1200, 1792 \end{array}\right\}$$

$$J_a = \{868, 1695, 569, 448, 1792\}$$

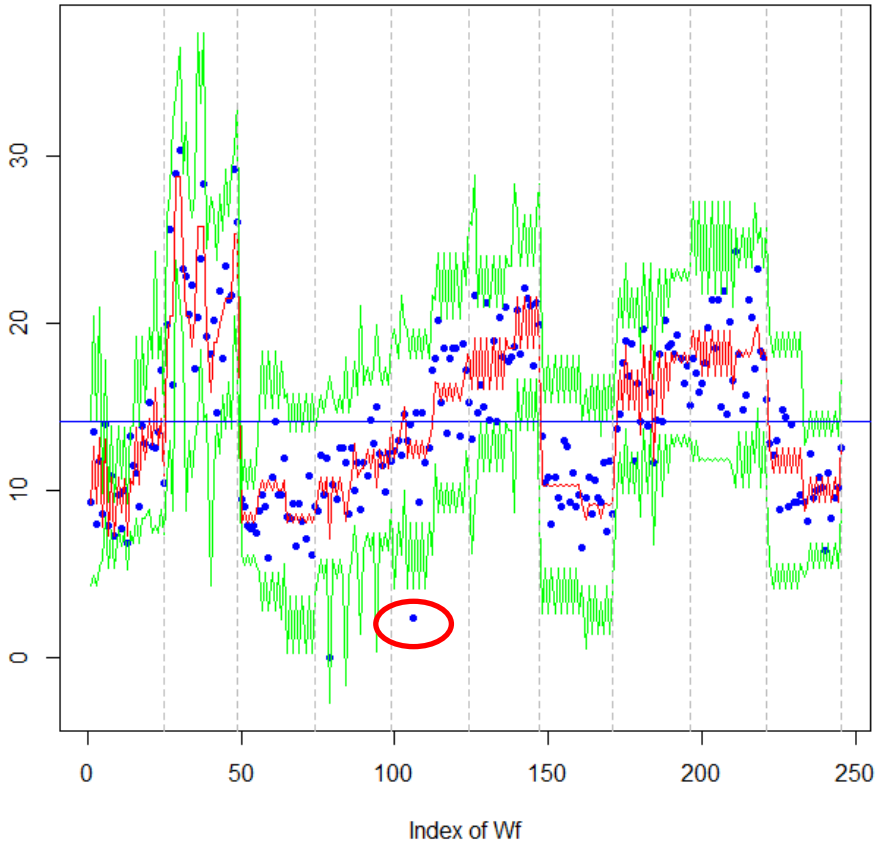
Case 1 配適模型(2015/09/10)

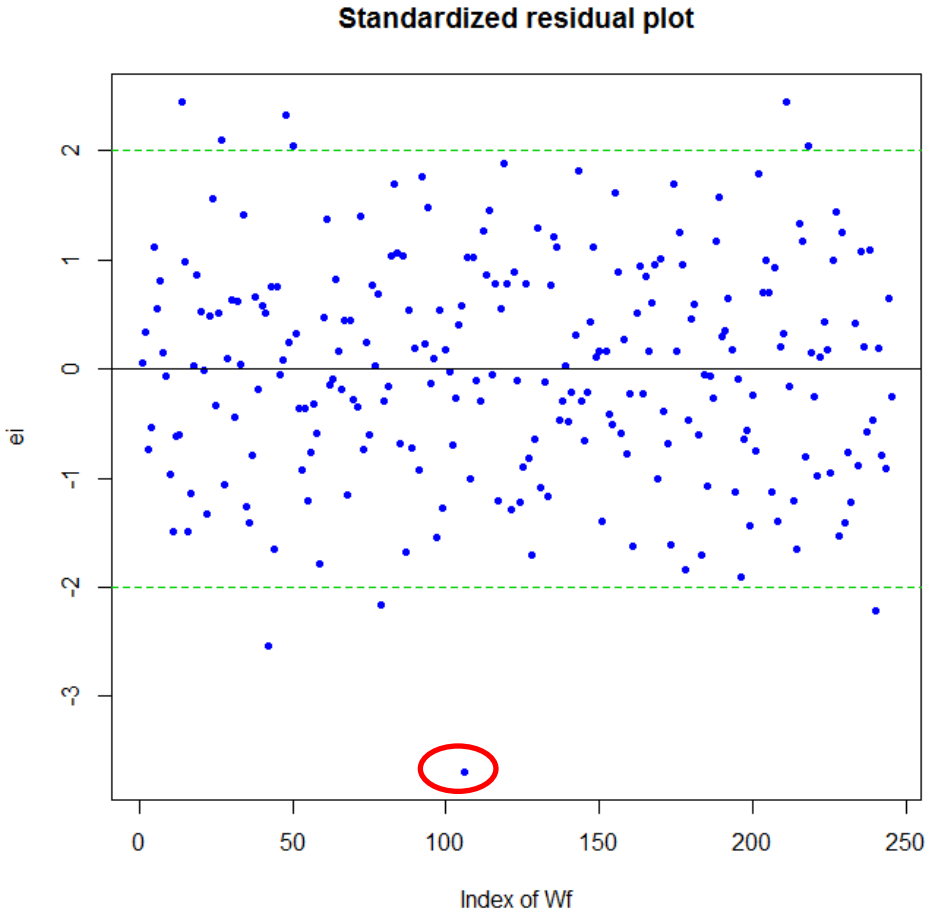
- 參數估計：





Case 1 分析表現(2015/09/10)





Case 1 重要機台(2015/09/10)

- Penalties: $wn0 = 0.25; wn1 = 0.55$
- 所挑選出機台結果：

$$J = \left\{ \begin{matrix} 404, 1370, 1586, 447, 1202, \underline{569}, 2542, 671, \\ 1, 446, \underline{868}, \underline{448}, \underline{1695} \end{matrix} \right\}$$

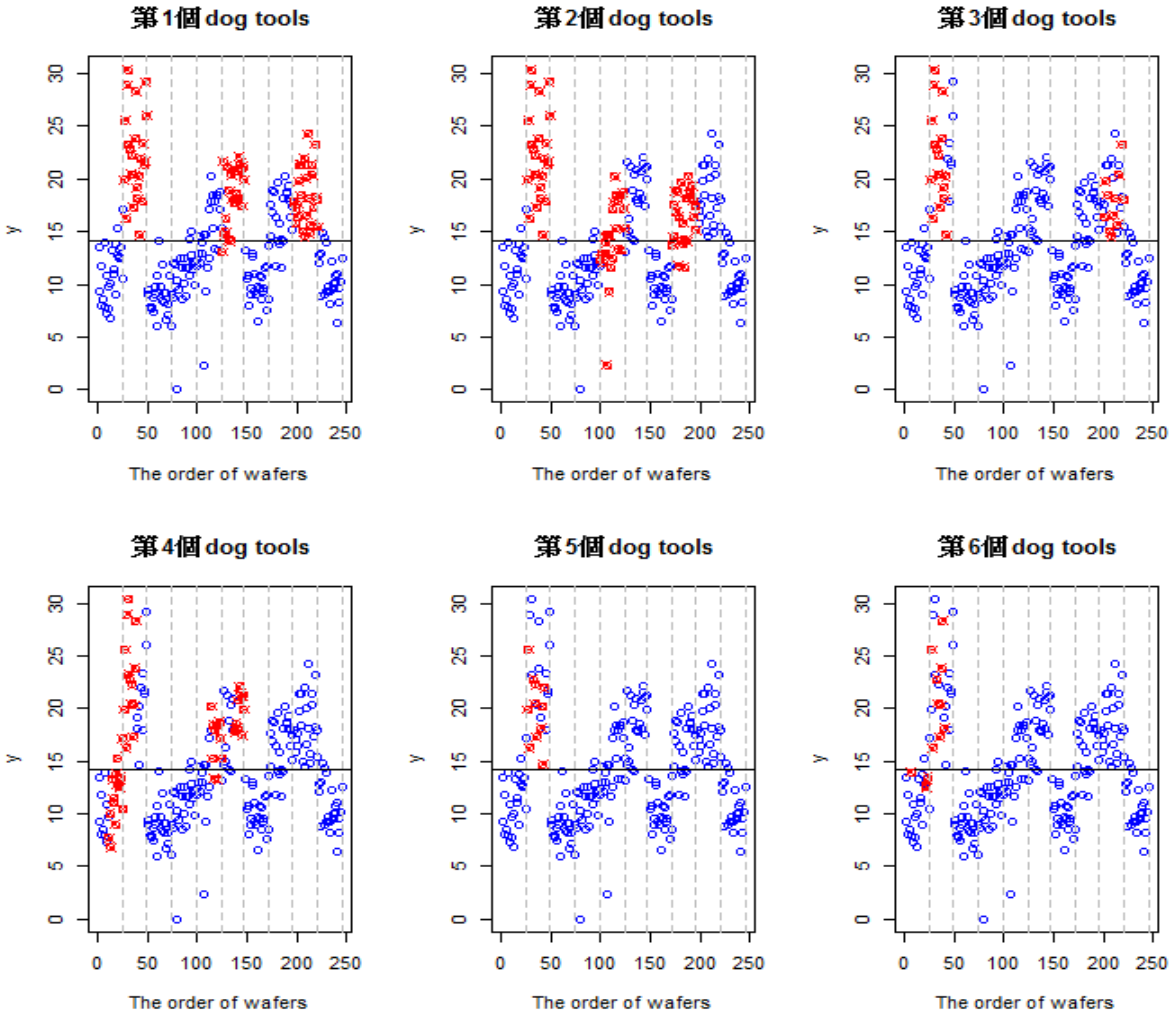
- 所挑選出之問題機台數：
  - Location: 10
  - Dispersion: 4

機台編號	機台名稱
404	S0106T0065
1026	S0265T0626
1222	S0315T0232
1490	S0375T0148

機台編號	機台名稱
1370	S0354T0048
2397	S0541T0293

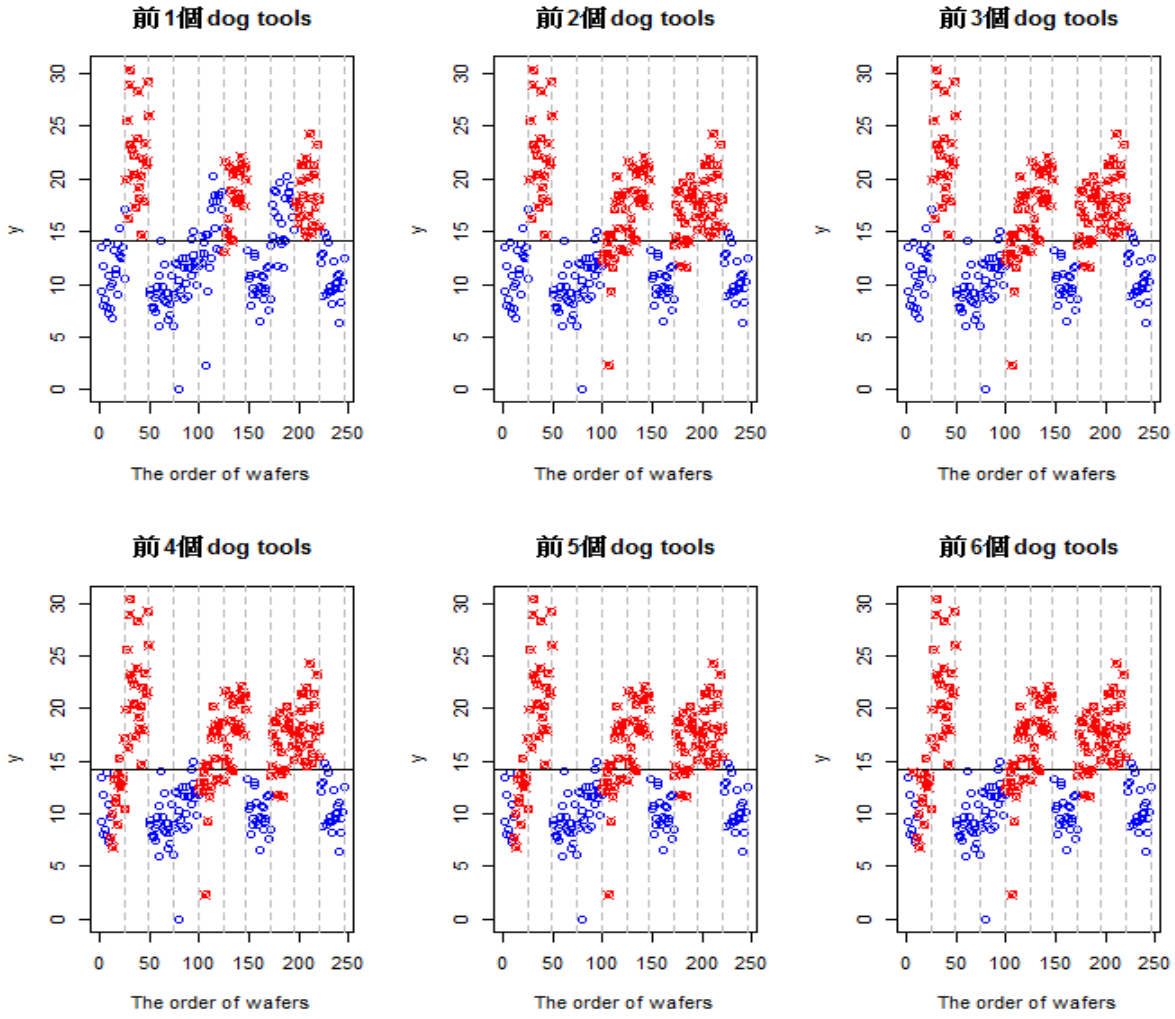
Case 1 分析表現(2015/09/10)

通過**第k個**重要機台的wafers表現



Case 1 分析表現(2015/09/10)

通過**前k個**重要機台的wafers表現



① 高维统计之现状与未来

- 成功的解决的高维线性回归问题
- 部分解决了高维矩阵估计问题.
- 有许多成功的应用
- 部分涉及非线性问题
- 扩大统计研究领域
- 高维及非参数的问题?
- 高维实时统计?
- 高维序贯方法? [e.g. multi-armed bandit problems with many covariates (contextual bandit problems)]
  - Alpha go (两个棋)
  - digital marketing
  - personalized medicine
- 高维 coordinate descent method ?

⋮

绪论：

关于统计模型选择

特点：
- 实用性高
- 易于产生跨领域合作
- 容易形成研究团队
- 容易吸引非数统背景学生

要求：
- 好的数学背景（逼近理论，分析…）
- 好的私事统计训练（大、小样本理论，参性及非参性模型之预测及推断、贝叶斯理论…）
- 好的计算能力（软件操作、程式，演算法分析…）
- 好的统计直觉？