The Covariance Inflation Criterion for Adaptive Model Selection

Robert Tibshirani and Keith Knight

Presented by
Arun Kumar, Mallik Rettiganti, Jie Ding and Dongmei Li

November 28, 2006

- Lets consider model selection problem in linear regression setting.
- We assume following linear model relates predictor x and response y.

$$Y = X\beta + \epsilon$$

  where X is $n \times p$ matrix with elements $x'_{ij}s$ and $\epsilon \sim N(0, I\sigma^2)$.
- We want to find out most important βs (out of p of them) using training set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where x_is are p-dimensional vectors.

- We will Use notation M_λ for best model with λ parameters. One can find such model using a variable selection procedure like forward stepwise regression for a fixed λ.
- We would like to select best model out of all $M'_\lambda s$ (there are p of them for $\lambda = 1, 2, 3, \ldots, p$) which minimizes prediction error:

$$PE(\hat{\beta}) = E||Y^* - X\hat{\beta}||^2$$

  Where Y^* comes from same distribution as y_i's in the training data or in other words $Y^* = X\beta + \epsilon^*$,$\epsilon^*$ is iid copy of ϵ.

- Can break prediction error in two parts like following:

$$\begin{aligned} PE(\hat{\beta}) &= E||Y^* - X\hat{\beta}||^2 \\ &= E||X\beta - X\hat{\beta} + \epsilon^*||^2 \\ &= E||X\beta - X\hat{\beta}||^2 + op(\lambda) \\ &= \bar{err}(\lambda) + op(\lambda) \end{aligned}$$

- If we know estimates of $\bar{err}(\lambda)$ and $op(\lambda)$ as function of λ then we can minimize sum of estimates to find λ and hence corresponding model M_λ will be our final model.

- We can estimate $\bar{err}(\lambda)$ by SSE(λ).
- Well known methods like Mallow's C_p and AIC(in general setting) estimates $op(\lambda)$ by $\frac{2\lambda\sigma^2}{n}$.
- This estimate of $op(\lambda)$ depends on λ only and does not include p (total number of possible parameters) which means p does not affect model selection procedure.
- Covariance inflation criteria proposed in this paper uses a different estimate of $op(\lambda)$. Authors claim that new estimate does depend on p and is better estimate of $op(\lambda)$.

- Proposed model selection criteria is based on
  $cic(\lambda) = \bar{err}(\lambda) + \hat{op}(\lambda)$
  where

$$\hat{op}(\lambda) = \frac{2\hat{\sigma}}{n\sigma_y^2} \sum_1^n cov^0\{y_i^*, \eta_{z^*}(x_i, M_\lambda^*)\} + \frac{2\hat{\sigma}^2}{n}$$

  σ_y^2 is sample variance for responses in training set and cov^0 indicates covariance between responses and predictions under the permutation distribution. We want smaller *cic*.
- Multiplicative factor $\frac{\hat{\sigma^2}}{\sigma_y^2}$ and last factor $\frac{2\hat{\sigma^2}}{n}$ are there to make estimate unbiased. Look at page 531-532 of the paper for proof!

- How is $\sum cov^0\{y_i^*, \eta_{z^*}(x_i, M_\lambda^*)\}$ calculated?
  - Keep x=$(x_1, \ldots, x_n)$ fixed.
  - Generate B random permutations of the responses $y^{*b} = \{y_1^{*b}, \ldots, y_n^{*b}\}, b = 1, 2, \ldots, B$.
  - Apply modeling procedure M_λ for each λ to the data set $\{(x_1, y_1^{*b}), \ldots, (x_2, y_2^{*b})\}$ and obtain fitted values η_i^{*b} for $i = 1, \ldots, n, \quad b = 1, 2, \ldots, B$.
  - Estimate covariance using following formulla

$$\sum_{i=1}^{n}\sum_{b=1}^{B}\frac{(y_i^{*b} - \bar{y})\eta_i^{*b}}{B}$$

- Why this make sense?
- More the number of covariates in the model,smaller $\bar{err}(\lambda)$ is.
- We would like to add a penalty for adding more variables though.
- As number of parameters increase, predictions for any permutation will be close to itself giving bigger covariance.
- Hence as we increase parameters in the model we reduce $\bar{err}(\lambda)$ at the cost of increasing $op(\lambda)$.

- Find M_λ for $\lambda = 1, 2, \ldots, p$ using model selection procedure(like forward stepwise regression).
- Calculate *cic* for each λ.
- Select λ which minimizes *cic*. Corresponding M_λ is the final model.

Orthogonal linear regression - The setup

- Consider the model $\mathbf{y} = \mathbf{X}\beta + \epsilon$, where $\mathbf{X}_{n\times p}$ has elements x_{ij} and $\mathbf{X}'\mathbf{X} = I$.
- Here λ is the number of predictors in the model.
- M_λ uses $\hat{\beta}_\lambda$, the LSE for the model corresponding to the best subset of size λ, i.e. the subset that gives the smallest residual sum of squares.
- This means, if t_j^2 is the squared t-statistic for the j^{th} predictor, then the best subset of size λ consists of the λ predictors having the largest value of t_j^2.

- Consider $y_i^* = \bar{y} + \epsilon^*$ where $\epsilon^* \sim N(0, \hat{\sigma}^2)$. This is asymptotically equivalent to permutation distribution.
- Then the correction term in the CIC becomes

$$\begin{aligned}\frac{2}{n}\frac{\hat{\sigma}^2}{\hat{\sigma}_y^2}\sum_{i=1}^{n}\mathrm{cov}^0\{y_i^*, \eta_{z^*}(x_i, M_\lambda)\} &= \frac{2}{n}E^0\left(\sum_{j=1}^{\lambda} t_{(j)}^2\right)\hat{\sigma}^2 \\ &\approx \frac{2}{n}\sum_{j=1}^{\lambda} 2\log\left(\frac{p}{j}\right).\end{aligned}$$

- For $\lambda = 1$, this becomes a simple threshold rule. Retain the predictor j if $t_{(1)}^2 > 4\log(p)$.
- This is very similar to RIC of Foster and George (1994). There it is shown that

$$E\{\max(t_j^2)\} \approx 2\log(p).$$

- For any λ, the CIC threshold is

$$\frac{4}{\lambda}\sum_{j=1}^{\lambda}\log\left(\frac{p}{j}\right).$$

- AIC and BIC correction terms (Schwarz, 1979) are also similar. Thus CIC can be compared with RIC, AIC and BIC.

Orthogonal Regression - Comparing CIC and RIC

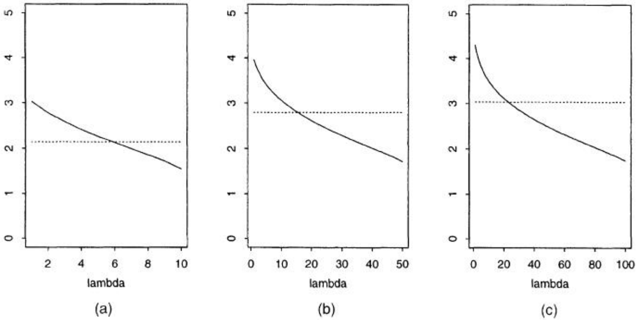

Fig. 1. Cumulative average of the square root of the CIC threshold $\{(4/\lambda)\ \Sigma_{j=1}^{\lambda} \log(p/j)\}^{1/2}$ (———) and square root of the RIC threshold $\{2\log(p)\}^{1/2}$ (··········), where λ is the subset size (these should be thought of as average thresholds for t-statistics): (a) $p = 10$; (b) $p = 50$; (c) $p = 100$

Orthogonal Regression - Null and non null models

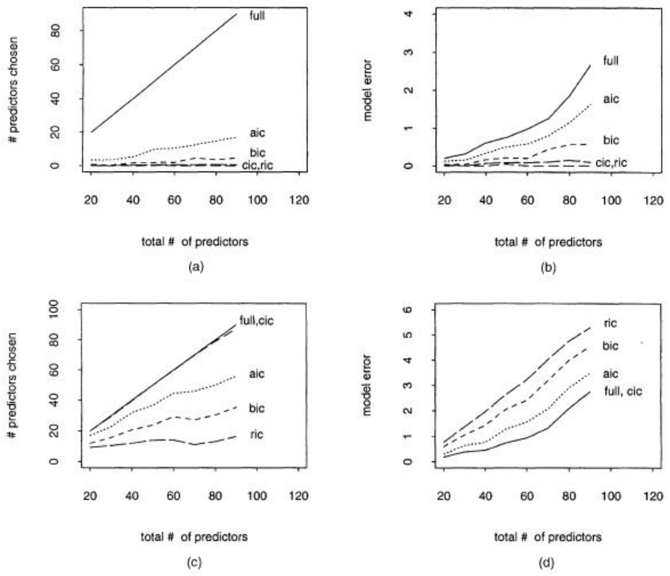

Fig. 2. (a), (b) Results for the null model and (c), (d) results for the non-null model, example 2 (the curves are means over five simulations; the standard errors of the means are about 1.5 on the left and 0.09 on the right)

Orthogonal Regression - Estimating true prediction error

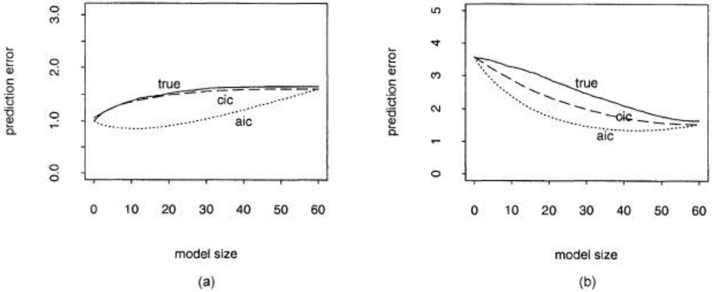

Fig. 3. Prediction error curves for (a) the null model and (b) the non-null model, example 2

- $p = 21$ predictors were used and $n = 50$ or $n = 150$ observations.
- $X_i's$ (predictors) are generated once according to multivariate normal with mean $\mathbf{0}$ and correlation $\text{corr}(X_j, X_k) = \rho^{|j-k|}$ with $\rho = 0.7$.
- The non-zero coefficients were generated in two clusters, around the 7$^{\text{th}}$ and the 14$^{\text{th}}$ predictors with initial values

$$\begin{aligned} \beta_{7+j} &= (h-j)^2, \qquad |j| < h \\ \beta_{14+j} &= (h-j)^2, \qquad |j| < h \end{aligned}$$

for $h = 1$ (a few strong effects), 2 (some moderate effects), 3 (many weak effects).

- In addition to these three, the null model scenario (all coefficients 0) and the full model scenario (all coefficients N(0,1)) were included in the study.
- $B = 10$ permutations were used to see if it would give satisfactory results.
- Interested in the size of the model selected $(\hat{\lambda})$, $ME(\lambda) = ||\hat{\mu}_\lambda - \mu||^2$ and its estimate $\hat{ME}(\hat{\lambda})$.
- Prediction Error $(PE) = ME + \sigma^2$. Since each criterion (AIC, BIC etc.) estimates PE, we can easily find $\hat{ME}$.

Table 1. Stepwise regression results†

Model	*Results for $n = 50$*						*Results for $n = 150$*					
	true	*oracle*	*cic*	*aic*	*cb*	*cv*	*true*	*oracle*	*cic*	*aic*	*cb*	*cv*
Null												
Size	0.00	0.00	0.00	2.60	5.40	0.00	0.00	0.00	0.00	3.20	0.60	0.00
ME	0.01	0.01	0.01	0.18	0.21	0.01	0.01	0.01	0.01	0.08	0.03	0.01
$\widehat{\mathrm{ME}}$		0.02	0.02	−0.04	−0.06	−0.18		−0.05	−0.05	−0.07	−0.05	−0.11
$h = 1$												
Size	2.00	2.00	2.20	5.40	6.40	2.20	2.00	2.00	2.40	5.80	5.60	2.40
ME	0.03	0.03	0.05	0.23	0.22	0.05	0.03	0.03	0.04	0.09	0.05	0.04
$\widehat{\mathrm{ME}}$		0.09	0.33	0.02	−0.01	0.07		−0.03	0.03	−0.05	−0.04	−0.04
$h = 2$												
Size	2.80	3.40	4.40	5.80	9.40	4.00	4.60	4.00	9.20	7.20	8.80	4.80
ME	0.15	0.22	0.26	0.31	0.32	0.24	0.05	0.06	0.11	0.09	0.10	0.09
$\widehat{\mathrm{ME}}$		0.28	0.39	0.03	0.06	0.23		0.05	0.08	−0.03	−0.01	0.01
$h = 3$												
Size	4.20	5.00	5.00	5.60	9.80	5.00	6.20	6.40	14.60	8.80	11.40	7.20
ME	0.22	0.29	0.29	0.32	0.33	0.36	0.09	0.11	0.13	0.12	0.13	0.11
$\widehat{\mathrm{ME}}$		0.38	0.41	0.04	0.06	0.32		0.08	0.09	−0.02	−0.02	0.05
Full												
Size	14.20	10.40	12.60	7.80	15.00	6.60	20.40	19.80	19.60	14.20	16.40	16.40
ME	0.32	0.43	0.45	0.44	0.43	0.55	0.14	0.15	0.14	0.18	0.17	0.18
$\widehat{\mathrm{ME}}$		0.46	0.50	0.16	0.07	0.55		0.08	0.09	0.02	−0.01	0.09
Average standard errors												
Size	0.61	0.69	1.10	0.86	2.82	0.82	0.30	0.18	1.55	0.97	2.28	0.80
ME	0.02	0.04	0.04	0.06	0.06	0.05	0.01	0.01	0.02	0.02	0.02	0.02
$\widehat{\mathrm{ME}}$		0.11	0.12	0.11	0.15	0.12		0.08	0.06	0.06	0.06	0.07

†Model size, actual model error ME and estimate of ME from each model, five settings: null model, $h = 1$ (a few strong effects), $h = 2$ (some moderate effects), $h = 3$ (many weak effects) and the full model. Methods: true, uses the actual ME; oracle, bootstrap samples from the true model to estimate optimism; cic, covariance inflation criterion; aic, Akaike's information criterion; cb, the conditional bootstrap; cv, tenfold cross-validation. The numbers are averages over 30 simulations. The last three rows give Monte Carlo standard errors.

1. For the null and $h = 1$ models, AIC chooses models that are too big and shows significant increase in ME.
2. For the $h = 2$ and $h = 3$ models, both AIC and CIC choose models that are too big, but the ME does not increase greatly.
3. For the full model, AIC underestimates the model size whereas CIC estimates it accurately.
4. AIC drastically underestimates the model error of its chosen model, whereas the CIC generally estimates it accurately.
5. For smaller sample size ($n = 50$), the conditional bootstrap overestimates the model size for the null, $h = 1$ and $h = 2$ and gives a poor estimate of the model error. For $n = 100$, it performs as good as the CIC.

General models

- Data: $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ with $z_i = (x_i, y_i)$ and $y_i \sim F_{\mu_i}$ independently
- Loss function: $Q[y, \eta]$
- Model: M_λ is a model of complexity λ chosen from the data

General models

- Data: $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ with $z_i = (x_i, y_i)$ and $y_i \sim F_{\mu_i}$ independently
- Loss function: $Q[y, \eta]$
- Model: M_λ is a model of complexity λ chosen from the data
- True error:

$$\mathrm{Err}(\lambda) = \frac{1}{n} \sum_{1}^{n} E_{\mu_i}\{Q[y_i^*, \eta_{\mathbf{z}}(x_i, M_\lambda)]\}$$

where $y_i^* \sim F_{\mu_i}$

General models

- Data: $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ with $z_i = (x_i, y_i)$ and $y_i \sim F_{\mu_i}$ independently
- Loss function: $Q[y, \eta]$
- Model: M_λ is a model of complexity λ chosen from the data
- True error:

$$\mathrm{Err}(\lambda) = \frac{1}{n} \sum_1^n E_{\mu_i} \{ Q[y_i^*, \eta_{\mathbf{z}}(x_i, M_\lambda)] \}$$

  where $y_i^* \sim F_{\mu_i}$
- Apparent error:

$$\overline{\mathrm{err}}(\lambda) = \frac{1}{n} \sum_1^n Q[y_i, \eta_{\mathbf{z}}(x_i, M_\lambda)]$$

Loss functions

- Commonly used loss function:

$$Q[y,\eta] = q(\eta) + \dot{q}(\eta)(y-\eta) - q(y)$$

where q is a concave function satisfying $q(0) = q(1) = 0$ and $\dot{q}$ is the derivative of q defined by left continuity

Loss functions

- Commonly used loss function:

$$Q[y, \eta] = q(\eta) + \dot{q}(\eta)(y - \eta) - q(y)$$

  where q is a concave function satisfying $q(0) = q(1) = 0$ and $\dot{q}$ is the derivative of q defined by left continuity
- Define $\hat{s} = -\frac{1}{2}\dot{q}(\eta)$

Loss functions

- Commonly used loss function:

$$Q[y,\eta] = q(\eta) + \dot{q}(\eta)(y-\eta) - q(y)$$

  where q is a concave function satisfying $q(0) = q(1) = 0$ and $\dot{q}$ is the derivative of q defined by left continuity
- Define $\hat{s} = -\frac{1}{2}\dot{q}(\eta)$
- Some common choices for Q

$Q[y,\eta]$	Possible values of y,η	$\hat{s}$
$(y-\eta)^2$	$y,\eta \in R$	η
$y\log(\eta) + (1-y)\log(1-\eta)$	$y = 0$ or 1, $\eta \in [0,1]$	$\log(\eta/(1-\eta))$
$I(y \neq \eta)$	$y,\eta = 0$ or 1	η

CIC for general models

- Define CIC:

$$\mathrm{cic}(\lambda) = \overline{\mathrm{err}(\lambda)} + \frac{2}{n}\sum_{1}^{n} \mathrm{Cov}^0(y_i, \hat{s}_i^*) + \frac{2}{n}$$

CIC for general models

- Define CIC:

$$\mathrm{cic}(\lambda) = \overline{\mathrm{err}(\lambda)} + \frac{2}{n}\sum_{1}^{n} \mathrm{Cov}^0(y_i, \hat{s}_i^*) + \frac{2}{n}$$

- For the loss functions satisfying the equation in previous slide, Efron(1986) proved

$$E\{\mathrm{Err}(\lambda) - \overline{\mathrm{err}}(\lambda)\} = \frac{2}{n}\sum_{1}^{n} \mathrm{cov}_{\mu_i}(y_i, \hat{s}_i)$$

Exponential families and logistic regression

- For fixed linear ML fit of λ in the exponential families, using approximations to get

$$\frac{2}{n}\sum_{1}^{n}\mathrm{cov}_{\mu_i}(y_i^*, \hat{s}_i^*) \approx \frac{2}{n}\sum_{1}^{n}\mathrm{cov}^0(y_i^*, \hat{s}_i^*) \approx \frac{2\lambda}{n}$$

Exponential families and logistic regression

- For fixed linear ML fit of λ in the exponential families, using approximations to get

$$\frac{2}{n}\sum_{1}^{n}\mathrm{cov}_{\mu_i}(y_i^*,\hat{s}_i^*) \approx \frac{2}{n}\sum_{1}^{n}\mathrm{cov}^0(y_i^*,\hat{s}_i^*) \approx \frac{2\lambda}{n}$$

- Use the set-up in section 4 and adapt it to logistic regression
- Define the binary response Y_i' by

$$\mathrm{Prob}(Y_i' = 1) = 1/(1+\exp(-\mu_i))$$

Exponential families and logistic regression

- For fixed linear ML fit of λ in the exponential families, using approximations to get

$$\frac{2}{n}\sum_{1}^{n} \mathrm{cov}_{\mu_i}(y_i^*, \hat{s}_i^*) \approx \frac{2}{n}\sum_{1}^{n} \mathrm{cov}^0(y_i^*, \hat{s}_i^*) \approx \frac{2\lambda}{n}$$

- Use the set-up in section 4 and adapt it to logistic regression
- Define the binary response Y_i' by

$$\mathrm{Prob}(Y_i' = 1) = 1/(1 + \exp(-\mu_i))$$

- For $n = 50$, AIC and conditional bootstrap chose models that are too big and had a large increase in prediction error while for $n = 150$ they did considerably better
- CIC and CV did well with CIC being better for smaller sample size while 10-fold CV tended to underestimate the model size for $n = 50$

One-nearest-neighbour classifier

- With two classes $y = 0$ and $y = 1$, let $\mu_i = \mathrm{Prob}(y_i = 1), \hat{\mu} = \frac{1}{n}\sum y_i$
- $$\frac{2}{n}\sum_{1}^{n} \mathrm{cov}_{\mu_i}(y_i^*, \hat{s}_i^*) = \frac{2}{n}\sum \mu_i(1 - \mu_i)$$
$$\frac{2}{n}\sum_{i}^{n} \mathrm{cov}^0(y_i^*, \hat{s}_i^*) + \frac{2}{n} = 2\hat{\mu}(1 - \hat{\mu}) + \frac{2}{n}$$

One-nearest-neighbour classifier

- With two classes $y = 0$ and $y = 1$, let $\mu_i = \mathrm{Prob}(y_i = 1), \hat{\mu} = \frac{1}{n}\sum y_i$
-
$$\frac{2}{n}\sum_{1}^{n} \mathrm{cov}_{\mu_i}(y_i^*, \hat{s}_i^*) = \frac{2}{n}\sum \mu_i(1-\mu_i)$$
$$\frac{2}{n}\sum_{i}^{n} \mathrm{cov}^0(y_i^*, \hat{s}_i^*) + \frac{2}{n} = 2\hat{\mu}(1-\hat{\mu}) + \frac{2}{n}$$
- By Jensen's inequality
$$E(2\hat{\mu}(1-\hat{\mu}) + \frac{2}{n}) > E(2\hat{\mu}(1-\hat{\mu})) \geq \sum \frac{2}{n}\mu_i(1-\mu_i)$$
- So CIC is biased upwards and will not work well for selecting the number of near neighbours.

Effective number of parameters

- When fitting a fixed linear model with λ parameters, the optimism is $\frac{2\lambda}{n}\sigma^2$

Effective number of parameters

- When fitting a fixed linear model with λ parameters, the optimism is $\frac{2\lambda}{n}\sigma^2$
- With prediction $\eta_i = \eta_{\mathbf{z}}(x_i, M_\lambda)$ the actual optimism is

$$\frac{2}{n}\sum_{1}^{n} \mathrm{cov}_{\mu_i}(y_i^*, \eta_i^*)$$

- The CIC estimate is

$$\frac{2}{n}\frac{\hat{\sigma}^2}{\sigma_y^2}\sum_{i}^{n} \mathrm{cov}^0(y_i^*, \eta_i^*) + \frac{2\hat{\sigma}^2}{n}$$

Effective number of parameters

- When fitting a fixed linear model with λ parameters, the optimism is $\frac{2\lambda}{n}\sigma^2$
- With prediction $\eta_i = \eta_{\mathbf{z}}(x_i, M_\lambda)$ the actual optimism is

$$\frac{2}{n}\sum_{1}^{n} \mathrm{cov}_{\mu_i}(y_i^*, \eta_i^*)$$

- The CIC estimate is

$$\frac{2}{n}\frac{\hat{\sigma}^2}{\sigma_y^2}\sum_{i}^{n} \mathrm{cov}^0(y_i^*, \eta_i^*) + \frac{2\hat{\sigma}^2}{n}$$

- Equate these with $2\lambda/n$ to get

$$\mathrm{enp}(\lambda) \equiv \sum \mathrm{cov}_{\mu_i}(y_i^*, \eta_i^*)$$

$$\widehat{\mathrm{enp}}(\lambda) \equiv \frac{1}{\sigma_y^2}\sum \mathrm{cov}^0(y_i^*, \eta_i^*) + 1$$

ENP for the orthogonal regression

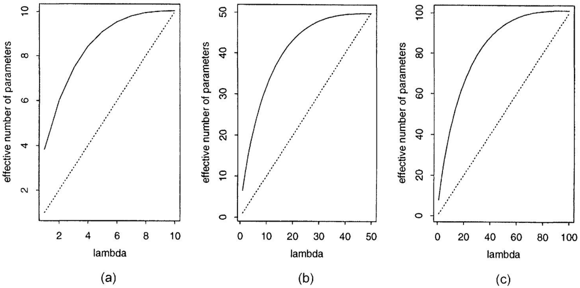

Fig. 4. Effective number of parameters (———) for the orthogonal all-subsets regression of example 2 and the 45°-line (··········) (λ is the subset size): (a) total number of predictors $p = 10$; (b) $p = 50$; (c) $p = 100$

- Adaptive selection makes the effective number of parameters greater than the nominal number of parameters λ, sometimes by a factor of 2

CIC for adaptive modelling procedure

- Adaptive modelling procedure for regression problem
  $y \longrightarrow M_\lambda \longrightarrow \hat{r} = H_\lambda y$

$$
\begin{aligned}
& \sum cov(y_i^\star, \eta_i^\star) \\
= \; & E_{M_\lambda}[tr\{H_\lambda\} \cdot var(y^\star|M_\lambda)] \\
& + E_{M_\lambda}(tr[H_\lambda \cdot \{E(y^\star|M_\lambda) - E(y^\star)\}\{E(y^\star|M_\lambda) - E(y^\star)\}]^T) \\
= \; & tr(H_\lambda)\sigma^2 + E_{M_\lambda}(tr[H_\lambda \cdot \{var(y^\star|M_\lambda) - var(y^\star)\}]) \\
& + E_{M_\lambda}(tr[H_\lambda \cdot \{E(y^\star|M_\lambda) - E(y^\star)\}\{E(y^\star|M_\lambda) - E(y^\star)\}]^T) \\
= \; & tr(H_\lambda)\sigma^2 + A(\lambda) + B(\lambda)
\end{aligned}
$$

CIC for adaptive modelling procedure

- Adaptive modelling procedure for regression problem
  $y \rightarrow M_\lambda \rightarrow \hat{r} = H_\lambda y$

$$
\begin{aligned}
&\sum cov(y_i^\star, \eta_i^\star) \\
= \; & E_{M_\lambda}[tr\{H_\lambda\} \cdot var(y^\star | M_\lambda)] \\
& + E_{M_\lambda}(tr[H_\lambda \cdot \{E(y^\star|M_\lambda) - E(y^\star)\}\{E(y^\star|M_\lambda) - E(y^\star)\}]^T) \\
= \; & tr(H_\lambda)\sigma^2 + E_{M_\lambda}(tr[H_\lambda \cdot \{var(y^\star|M_\lambda) - var(y^\star)\}]) \\
& + E_{M_\lambda}(tr[H_\lambda \cdot \{E(y^\star|M_\lambda) - E(y^\star)\}\{E(y^\star|M_\lambda) - E(y^\star)\}]^T) \\
= \; & tr(H_\lambda)\sigma^2 + A(\lambda) + B(\lambda)
\end{aligned}
$$

- $tr(H_\lambda)\sigma^2$ is the non-adaptive part of the error, $A(\lambda)$ and $B(\lambda)$ capture the adaptive component.
  $A(\lambda)$ and $B(\lambda)$ are 0 under a fixed model choice.

Properties of CIC

- The CIC estimate of the prediction error curve is biased unless the true model is null.

Properties of CIC

- The CIC estimate of the prediction error curve is biased unless the true model is null.
- CIC seems overestimate the optimism when $\lambda < \lambda_0$ and roughly unbiased for the optimism when $\lambda \geq \lambda_0$ from the simulation results.

Properties of CIC (Continued)

- The CIC is not a consistent model selection method in the sense of choosing the smallest 'correct' model with probability tending to 1.

Properties of CIC (Continued)

- The CIC is not a consistent model selection method in the sense of choosing the smallest 'correct' model with probability tending to 1.
- When $\lambda > \lambda_0$
  $cic(\lambda) = \overline{err}(\lambda) + \widehat{op}(\lambda)$
  $n\{\overline{err}(\lambda_0) - \overline{err}(\lambda)\} \xrightarrow{d} \sigma^2 \chi^2_l$
  $n\{\widehat{op}(\lambda) - \widehat{op}(\lambda_0)\} \leq Ml$ for some M $\Rightarrow cic(\lambda) < cic(\lambda_0)$ with positive probability.

Properties of CIC (Continued)

- The CIC is not a consistent model selection method in the sense of choosing the smallest 'correct' model with probability tending to 1.
- When $\lambda > \lambda_0$
  $cic(\lambda) = \overline{err}(\lambda) + \widehat{op}(\lambda)$
  $n\{\overline{err}(\lambda_0) - \overline{err}(\lambda)\} \xrightarrow{d} \sigma^2\chi^2_l$
  $n\{\widehat{op}(\lambda) - \widehat{op}(\lambda_0)\} \leq Ml$ for some M $\Rightarrow cic(\lambda) < cic(\lambda_0)$
  with positive probability.
- When $\lambda < \lambda_0$
  $\overline{err}(\lambda) - \overline{err}(\lambda_0) \xrightarrow{p} \gamma > 0$
  $\widehat{op}(\lambda) - \widehat{op}(\lambda_0) \xrightarrow{p} 0 \Rightarrow P\{cic(\lambda) < cic(\lambda_0)\} \rightarrow 0$

Discussion

- The CIC is useful for a variety of adaptive fitting methods although it works poorly in extremely overfitted methods.

Discussion

- The CIC is useful for a variety of adaptive fitting methods although it works poorly in extremely overfitted methods.
- Little bootstrap procedure of Breiman (1992)(similar to CIC):
  $y_i^\star = y_i + \epsilon_i$, $\epsilon_i \sim N(0, t^2 sig\hat{m}a^2)$
  $\hat{op} = (1/t^2) \sum cov(\hat{y}_i{}^\star, \epsilon_i)/n$

Discussion

- The CIC is useful for a variety of adaptive fitting methods although it works poorly in extremely overfitted methods.
- Little bootstrap procedure of Breiman (1992)(similar to CIC):
  $y_i^\star = y_i + \epsilon_i$, $\epsilon_i \sim N(0, t^2 \hat{sigma}^2)$
  $\hat{op} = (1/t^2) \sum cov(\hat{y}_i{}^\star, \epsilon_i)/n$
- For small t, above estimator is an approximately unbiased estimate of the optimism, however its variance becomes large when $t \rightarrow 0$. $t = 0.6$ was recommended from empirical studies.

Discussion (Cont'd)

- Cross-validation performs well in simulation.

Discussion (Cont'd)

- Cross-validation performs well in simulation.
- Ordinary cross-validation estimate of the prediction error is: $\widehat{Err}^{(cvl)}(\lambda) = \frac{1}{n}\sum_1^n Q[y_i, \eta_{Z_i}(X, \lambda)]$

Discussion (Cont'd)

- Cross-validation performs well in simulation.
- Ordinary cross-validation estimate of the prediction error is: $\widehat{Err}^{(cvl)}(\lambda) = \frac{1}{n}\sum_1^n Q[y_i, \eta_{Z_i}(X, \lambda)]$
- Cross-validation estimates "extra-sample error": $Err_{ex}(z, \lambda) = E_{0G} Q[y_0, \eta_z(x_0, M_\lambda)]$
  $(x_0, y_0) \sim G$ with z held fixed.

Discussion (Cont'd)

- Cross-validation performs well in simulation.
- Ordinary cross-validation estimate of the prediction error is: $\widehat{Err}^{(cvl)}(\lambda) = \frac{1}{n}\sum_1^n Q[y_i, \eta_{Z_i}(X, \lambda)]$
- Cross-validation estimates "extra-sample error": $Err_{ex}(z, \lambda) = E_{0G} Q[y_0, \eta_z(x_0, M_\lambda)]$ $(x_0, y_0) \sim G$ with z held fixed.
- Bootstrap estimates can be seen as estimates of extra-sample error.

Discussion (Cont'd)

- Cross-validation performs well in simulation.
- Ordinary cross-validation estimate of the prediction error is: $\widehat{Err}^{(cvl)}(\lambda) = \frac{1}{n}\sum_1^n Q[y_i, \eta_{Z_i}(X, \lambda)]$
- Cross-validation estimates "extra-sample error": $Err_{ex}(z, \lambda) = E_{0G} Q[y_0, \eta_z(x_0, M_\lambda)]$ $(x_0, y_0) \sim G$ with z held fixed.
- Bootstrap estimates can be seen as estimates of extra-sample error.
- Cross-validation can cause bias when using a small training set in each fold.

Discussion (Cont'd)

- A loose relationship between CIC and structure risk minimization (Vapnik, 1996).
  $E(Err) \leq E(\overline{err}) + f(h, n)$
  h is the Vapnik-Chervonenkis (VC) dimension of the model and f is a known function.
  Similar to CIC in permutation operation, but does not capture the effect of adaptive fitting.

Discussion (Cont'd)

- A loose relationship between CIC and structure risk minimization (Vapnik, 1996).
  $E(Err) \leq E(\overline{err}) + f(h, n)$
  h is the Vapnik-Chervonenkis (VC) dimension of the model and f is a known function.
  Similar to CIC in permutation operation, but does not capture the effect of adaptive fitting.
- CIC estimates the true optimism by using permutation distribution of x and y. The reason that permutation works is that it is a good estimate of marginal distribution of y, which is the quantity of interest for linear estimators.

Discussion (Cont'd)

- A loose relationship between CIC and structure risk minimization (Vapnik, 1996).
  $E(Err) \leq E(\overline{err}) + f(h, n)$
  h is the Vapnik-Chervonenkis (VC) dimension of the model and f is a known function.
  Similar to CIC in permutation operation, but does not capture the effect of adaptive fitting.
- CIC estimates the true optimism by using permutation distribution of x and y. The reason that permutation works is that it is a good estimate of marginal distribution of y, which is the quantity of interest for linear estimators.
- Using null bootstrap distribution can also gives unbiased estimate of covariance of $\hat{\beta}$