A Prediction Divergence Criterion for Model Selection

Stéphane Guerrier (UCSB)

joint work with

Maria-Pia Victoria-Feser (University of Geneva)

12^{th} of February, 2014

Introduction

What is model selection?

- Model selection is a crucial part of any statistical analysis.
- Inevitable in an increasingly large number of applications involving partial theoretical knowledge and vast amounts of information, like in medicine, biology or economics.
- Model selection can be **applied to any situation where one tries to balance variability with complexity**.
- For example, these techniques can be applied to **select "significant" variables in regression problems**, to determine the number of dimensions in principal component analyses or simply to construct histograms.

Introduction

Most common variable selection strategies are:

- **Tests** (e.g. t-test, goodness-of-fit test, ...):
  - \+ Several computationally efficient approaches for with large datasets (e.g. *Foster & Stine, 2004*).
  - \- Tends to poorly perform in terms of prediction error.
- **Nonparametric "bootstrap" techniques** (e.g. cross-validation, ...)
  - \- Pays a substantial price in terms of decreased estimating efficiency (*Efron, 2004*).
- **Model selection criterion** (e.g. AIC, BIC, ...):
  - \+ Computational speed when used as a sequential method.
  - \- Instability (*Breiman, 1996*).
- **Regularisation** (e.g. lasso, Dantzig selector, ...):
  - \+ Increasingly being used and often provide better results than other selection approaches.

Introduction

- Suppose that we have two (nested) models, say $\mathcal{M}_1$ and $\mathcal{M}_2$ (associated to the predictions $\hat{\boldsymbol{y}}_1$ and $\hat{\boldsymbol{y}}_2$).
- Assume that we consider a model selection criterion C (say the AIC) to determine whether $\mathcal{M}_1$ or $\mathcal{M}_2$ is more suitable to predict (future realisation of) $\boldsymbol{y}$ (i.e. $\boldsymbol{y}^0$).
- We would based our decision on $\Delta\,\text{AIC} = \text{AIC}_2 - \text{AIC}_1 = \mathbb{E}[\mathbb{E}_0[D(\boldsymbol{y}^0, \hat{\boldsymbol{y}}_2)]] - \mathbb{E}[\mathbb{E}_0[D(\boldsymbol{y}^0, \hat{\boldsymbol{y}}_1)]]$.
- **Initial idea**: consider (instead of $\Delta\,\text{AIC}$) a quantity of the form $\text{C} = \mathbb{E}[\mathbb{E}_0[D(\hat{\boldsymbol{y}}_1^0, \hat{\boldsymbol{y}}_2)]]$.
- **Intuitively**: $\mathbb{E}[\mathbb{E}_0[D(\hat{\boldsymbol{y}}_1^0, \hat{\boldsymbol{y}}_2)]]$ is small (large) $\rightarrow$ select $\mathcal{M}_1$ ($\mathcal{M}_2$).
- **We will see that this approach performs extremely well in "sparse" settings.**
- Consider a simple example with $\mathbb{E}[\mathbb{E}_0[D(\hat{\boldsymbol{y}}_1^0, \hat{\boldsymbol{y}}_2)]] = \mathbb{E}[\mathbb{E}_0[||\hat{\boldsymbol{y}}_1^0 - \hat{\boldsymbol{y}}_2||_2^2]]$.

Motivating Example

Setting:

- We consider linear models: $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma_{\varepsilon}^2 \mathbf{I}\right)$ and where the matrix $\mathbf{X}$ is randomly generating for each replication such that $\mathbf{x}_i \sim \mathcal{N}\left(\mathbf{0}_p, \boldsymbol{\Sigma}\right)$.
- **Model**:

$$\boldsymbol{\beta} = (0, 1, 0, 1, 0, 1, 0, 1, 0, 1, \underbrace{0, ..., 0}_{50})$$

with $\sigma_{\epsilon}^2 = 1$, $\text{corr}(\mathbf{x}_j, \mathbf{x}_k) = 0.5^{|j-k|}$ and $n = 80$.

- Comparison with:
  - FPE (*Akaike, 1969*), AIC (*Akaike, 1974*), AICc (*Hurvich and Tsai, 1989*), FPEu and AICu (*McQuarrie et al., 1997*).
  - BIC (*Schwarz (1978)*), HQ (*Hannan and Quinn (1979)*) and HQc (*McQuarrie and Tsai (1998)*).
  - Lasso (*Tibshirani, 1996*) and Elastic Net (*Zou & Hastie, 2005*).

Motivating Example

Setting:

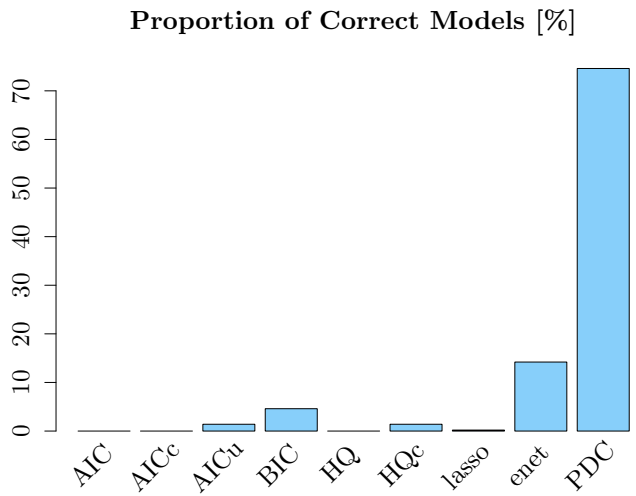
Criteria	Description
Cor. [%]	Proportion of times the correct model is selected.
Inc. [%]	Proportion of times the correct model is nested.
true+	Average number of selected significant variables.
false+	Average number of selected non-significant variables.
NbReg	Average number of regressors in the selected model.
Med $(\text{PE}_{\boldsymbol{y}})$	Median of $\text{PE}_{\boldsymbol{y}}$ computed on test samples.
Med $(\text{MSE}_{\boldsymbol{\beta}})$	Median of $\text{MSE}_{\boldsymbol{\beta}}$ computed on test samples.

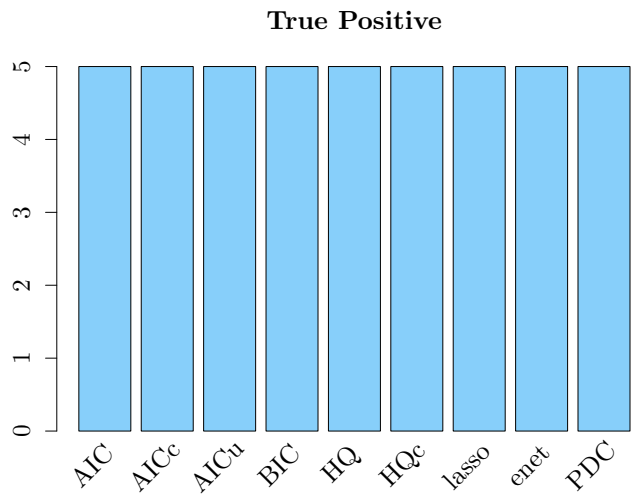
where

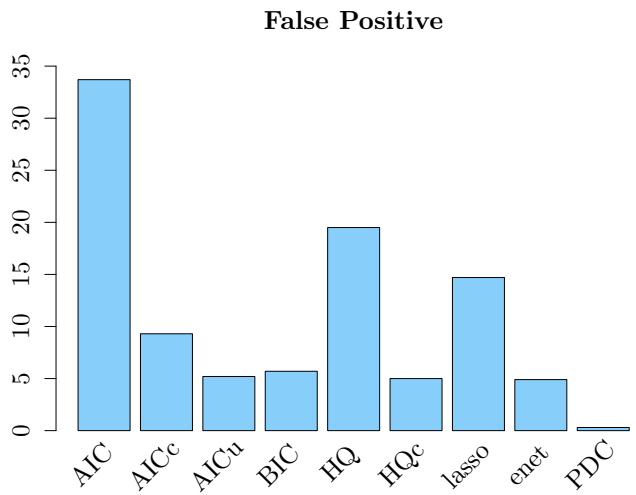
$$\text{PE}_{\boldsymbol{y}} = \frac{1}{n^{\star}} ||\boldsymbol{y}_{test} - \mathbf{X}_{test}\hat{\boldsymbol{\beta}}||_2^2$$

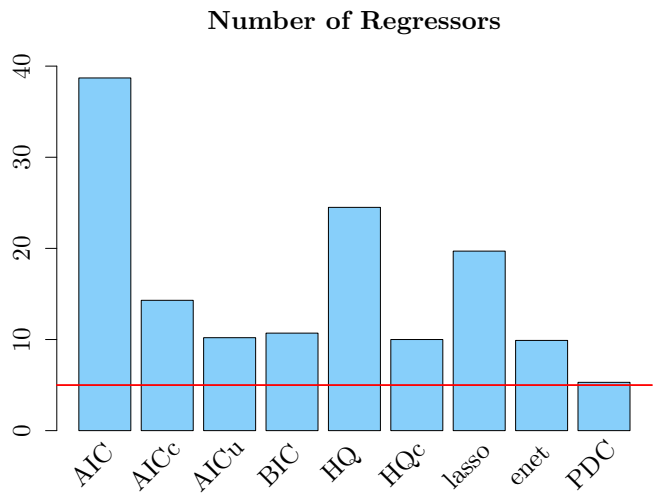
$$\text{MSE}_{\boldsymbol{\beta}} = ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2^2 .$$

Motivating Example

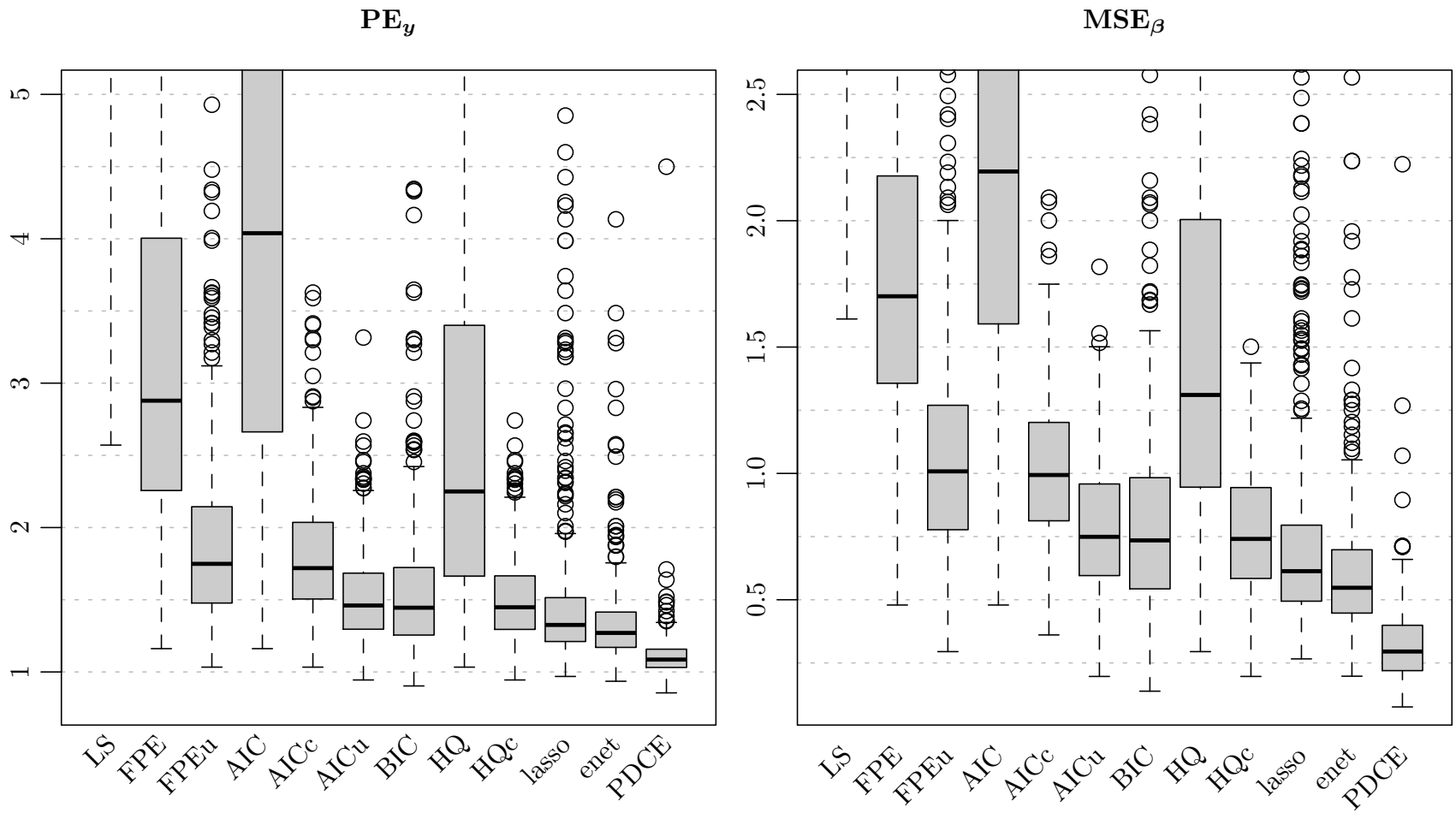








Motivating Example



Outline

1. Efron's covariance penalty criterion
2. The d-class of error measure
3. Optimism theorem for the d-class
4. The class of Prediction Divergence Criterion (PDC)
5. Application to linear models
6. Asymptotic properties:
   - Limiting distribution
   - Efficiency and consistency
7. Simulations
8. Extension(s): Generalized Linear Models
9. Application: Acute Leukemia Class Prediction

Model Selection

Setting:

- Consider a random variable Y distributed according to model $F_{\boldsymbol{\theta}}$, possibly conditionally on a set of fixed covariates $\mathbf{x} = [x_1 \dots x_p]$.
- We observe a random sample $\boldsymbol{Y} = (Y_i)_{i=1,\dots,n}$ supposedly generated from $F_{\boldsymbol{\theta}}$, possibly together with a non-random $n \times p$ full rank matrix of inputs $\mathbf{X}$.
- Consider a prediction function $\hat{\boldsymbol{Y}}$ that depends on the chosen model.
- We wish to determine the model(s) that best predicts future realisation of Y (notation: Y^0).

Efron's Covariance Penalty Criterion

Optimism theorem for q-class error measure:

Consider a q-**class error measure** (*Efron, 1987*):

$$Q(u, v) = q(v) + \dot{q}(v)(u - v) - q(u).$$

Example: $Q(u, v) = (u - v)^2 \ \rightarrow \ q(u) = u(1 - u)$.

Efron, 2004 derived the following result (under some regularity conditions):

$$\mathbb{E}\left[\mathbb{E}_0\left[Q\left(y_i^0, \hat{y}_i\right)\right]\right] = \mathbb{E}\left[Q\left(y_i, \hat{y}_i\right) + \mathrm{cov}\left(\dot{q}(\hat{y}_i), y_i\right)\right].$$

This result enables to derive in a unified manner, among others, the AIC and Mallow's C_p. The AIC and C_p correspond respectively to $Q(\cdot)$ being squared loss function and the Kullback-Leibler divergence.

The d-class Error Measure

Definition:

- Consider a **Bregman divergence**:

$$D(\mathbf{u}, \mathbf{v}) = \psi(\mathbf{u}) - \psi(\mathbf{v}) - (\mathbf{u} - \mathbf{v})^T \nabla\psi(\mathbf{v}).$$

- **Example**: $D(\mathbf{u}, \mathbf{v}) = ||\mathbf{u} - \mathbf{v}||_2^2 \rightarrow \psi(\mathbf{u}) = \mathbf{u}^T\mathbf{u}$.
- Consider two equidimensional vector valued functions $\boldsymbol{f}_1(\hat{\theta}_1, \boldsymbol{y})$ and $\boldsymbol{f}_2(\hat{\theta}_2, \boldsymbol{y})$.
- Let $\hat{\theta}_1$ and $\hat{\theta}_2$ denote the estimated parameter vectors associated, respectively, to the models $F_{\boldsymbol{\theta_1}}$ and $F_{\boldsymbol{\theta_2}}$.
- Then $D(\boldsymbol{f}_1(\boldsymbol{y}, \hat{\theta}_1), \boldsymbol{f}_2(\boldsymbol{y}, \hat{\theta}_2))$ is said to belong to the d-**class of error measures** (+ additional technical conditions).

Optimism Theorem

Theorem 1:

For a valid d-class of error measures (and under some regularity conditions) we have that

$$\mathrm{C} = \mathbb{E}\left[\mathbb{E}_0\left[D\left(\boldsymbol{f}_1\left(\boldsymbol{y}^0, \hat{\boldsymbol{\theta}}_1\right), \boldsymbol{f}_2\left(\boldsymbol{y}, \hat{\boldsymbol{\theta}}_2\right)\right)\right]\right] =$$

$$\mathbb{E}\left[D\left(\boldsymbol{f}_1\left(\boldsymbol{y}, \hat{\boldsymbol{\theta}}_1\right), \boldsymbol{f}_2\left(\boldsymbol{y}, \hat{\boldsymbol{\theta}}_2\right)\right)\right] + \mathrm{tr}\left\{\mathrm{cov}\left[\boldsymbol{f}_1\left(\boldsymbol{y}, \hat{\boldsymbol{\theta}}_1\right), \nabla\psi\left(\boldsymbol{f}_2\left(\boldsymbol{y}, \hat{\boldsymbol{\theta}}_2\right)\right)\right]\right\}.$$

This result can be used to derive the AIC, C_p and other common criteria. It also enables to derive model selection criteria for GMM, robust estimators, etc.

Remark:

A "natural" (and consistent) estimator of C is

$$\widehat{\mathrm{C}} = D\left(\boldsymbol{f}_1\left(\boldsymbol{y}, \hat{\boldsymbol{\theta}}_1\right), \boldsymbol{f}_2\left(\boldsymbol{y}, \hat{\boldsymbol{\theta}}_2\right)\right) + \mathrm{tr}\left\{\widehat{\mathrm{cov}}\left[\boldsymbol{f}_1\left(\boldsymbol{y}, \hat{\boldsymbol{\theta}}_1\right), \nabla\psi\left(\boldsymbol{f}_2\left(\boldsymbol{y}, \hat{\boldsymbol{\theta}}_2\right)\right)\right]\right\}$$

where $\widehat{\mathrm{cov}}(\cdot)$ is obtained analytically up to a value of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ which are then replaced by $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$, or by resampling methods (see e.g. *Efron, 2004*).

Applications of the d-Class of Error Measure

1. "Classical" criteria ($C_p \to L_2$ loss, AIC $\to$ KL divergence, ...).
2. Prediction Divergence Criteria.
3. Robust model selection criteria (typically Mahalanobis loss).
4. Generalized Method of Wavelet Moments (Mahalanobis loss + Wavelet Variance).
5. ...

The Prediction Divergence Criterion

- When a model selection criterion, say C, is used in practice to choose between two models, say $\mathcal{M}_j$ nested in $\mathcal{M}_k$, an estimate of C is computed for both models and **the difference is used for selection**.
- We propose instead another class of criteria that aims at **directly measuring a prediction divergence between the two (nested) models**, i.e.

$$\text{PDC}_{j,k} = \mathbb{E}\left[\mathbb{E}_0\left[D\left(\hat{\boldsymbol{y}}_j^0, \hat{\boldsymbol{y}}_k\right)\right]\right].$$

- Using Theorem 1 we have that:

$$\widehat{\text{PDC}}_{j,k} = D\left(\hat{\boldsymbol{y}}_j, \hat{\boldsymbol{y}}_k\right) + \text{tr}\left\{\widehat{\text{cov}}\left[\hat{\boldsymbol{y}}_j, \nabla\psi\left(\hat{\boldsymbol{y}}_k\right)\right]\right\}.$$

The Prediction Divergence Criterion

- If $D(\cdot,\cdot)$ is the squared loss function, then $\nabla\psi(\mathbf{x}) = 2\mathbf{x}$, so we may write

$$\widehat{\text{PDC}}_{j,k} = ||\hat{\boldsymbol{y}}_j - \hat{\boldsymbol{y}}_k||_2^2 + 2\,\text{tr}\left[\widehat{\text{cov}}\left(\hat{\boldsymbol{y}}_j, \hat{\boldsymbol{y}}_k\right)\right].$$

- Similarly to AIC_α (see e.g. *Bhansali and Downham, 1977*) we defined $\widehat{\text{PDC}}^{\lambda_n}_{j,k}$ as:

$$\widehat{\text{PDC}}^{\lambda_n}_{j,k} = ||\hat{\boldsymbol{y}}_j - \hat{\boldsymbol{y}}_k||_2^2 + \lambda_n\,\text{tr}\left[\widehat{\text{cov}}\left(\hat{\boldsymbol{y}}_j, \hat{\boldsymbol{y}}_k\right)\right].$$

- In general the term $\widehat{\text{cov}}\left(\hat{\boldsymbol{y}}_j, \hat{\boldsymbol{y}}_k\right)$ is difficult to obtain but may be obtained by parametric bootstrap.

The Prediction Divergence Criterion

- For linear models we have:

$$\widehat{\mathrm{PDC}}_{j,k}^{\lambda_n} = ||\hat{\boldsymbol{y}}_j - \hat{\boldsymbol{y}}_k||_2^2 + \lambda_n \sigma_\varepsilon^2 \,\mathrm{tr}\,(\mathbf{S}_j \mathbf{S}_k) = ||\hat{\boldsymbol{y}}_j - \hat{\boldsymbol{y}}_k||_2^2 + \lambda_n \sigma_\varepsilon^2 j.$$

- Assuming that there exist K competing nested models, we propose to choose the model $\hat{\jmath}_{\lambda_n}$ satisfying:

$$\hat{\jmath}_{\lambda_n} = \underset{j=1,\ldots,K-1}{\mathrm{argmin}} \; \widehat{\mathrm{PDC}}_{j,j+1}^{\lambda_n}.$$

  This selection rule is motivated by Theorem 2.

- If a clear sequence of competing nested models does not exist, one can build one prior to applying the selection rule (see Theorem 6).

Asymptotic Properties: The Ordered Case

Simplified regression setting:

Consider

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n\right)$$

where $\boldsymbol{y}$ is a vector of observations, $\mathbf{X}$ is a full-rank $n \times K$ design matrix and $\boldsymbol{\beta}_j = [\beta_1, ..., \beta_j, 0, ..., 0]$ is a vector of regression parameters. We only consider here the problem of selecting the number j.

Let the class of models $\mathscr{J}$ correspond to all possible models. Suppose that the observations were generated by the following model $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$. Then we define the "best" model $\jmath_0 \in \mathscr{J}$ as

$$\jmath_0 = \max \left\{ j \in \{1, ..., K\} \; : \; \operatorname*{plim}_{n \to \infty} \left| \hat{\beta}_j^\star \right| > 0 \right\}$$

where $\hat{\beta}_j^\star$ denotes the j^{th} element of $\hat{\boldsymbol{\beta}}_K$.

Asymptotic Properties: The Ordered Case

Theorem 2:

Under some regularity conditions on $\hat{\sigma}_\varepsilon^2$ and λ_n as well as for j and m such that $0 < j < K$, $m > 0$ and $j + m \leq K + 1$ we have that for sufficiently large n

$$\mathbb{E}\left[\widehat{\text{PDC}}^{\lambda_n}_{\jmath_0,\jmath_0+1}\right] \leq \mathbb{E}\left[\widehat{\text{PDC}}^{\lambda_n}_{j,j+m}\right].$$

We also have that $\mathbb{E}\left[\widehat{\text{PDC}}^{\lambda_n}_{\jmath_0,\jmath_0+1}\right] = \mathbb{E}\left[\widehat{\text{PDC}}^{\lambda_n}_{j,j+m}\right]$ if and only if $j = \jmath_0$ and $m = 1$.

Remember that:

$$\hat{\jmath}_{\lambda_n} = \underset{0\leq j\leq K}{\text{argmin}}\ \widehat{\text{PDC}}^{\lambda_n}_{j,j+1} \quad \text{where} \quad \widehat{\text{PDC}}^{\lambda_n}_{j,j+1} = \left\|\hat{\boldsymbol{y}}_j - \hat{\boldsymbol{y}}_{j+1}\right\|_2^2 + \lambda_n j \hat{\sigma}_\varepsilon^2.$$

Assumptions

Suppose that the observations were generated by the following model:

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}.$$

Then we assume that:

(A1) When σ_ε^2 is unknown it can be replace by a consistent estimator $\hat{\sigma}_\varepsilon^2$.

(A2) The vector $\boldsymbol{\mu}$ is such that:

- $\boldsymbol{\mu}^T (\mathbf{S}_{j+1} - \mathbf{S}_j) \boldsymbol{\mu} = \mathcal{O}(n)$ if $j \notin \mathcal{J}_0$.
- $\boldsymbol{\mu}^T (\mathbf{S}_j - \mathbf{S}_{\mathcal{J}_0}) \boldsymbol{\mu} = \mathcal{O}(n)$ if $j \in \mathcal{J}_0$.

(A3) The scalars λ_n and $\jmath_0$ are such that $\lambda_n \jmath_0 = \mathcal{O}\left(\sqrt{n}\right)$ and $\lambda_n > 0$.

Remark:

Assumption (A1) can be replaced by:

(A4) The vector $\boldsymbol{\mu}$ is such that $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}_{\mathcal{J}_0}$.

Asymptotic Probability of Underfitting

Theorem 3:

Let $j \notin \mathscr{J}_0$, then under Assumptions (A1), (A2) and (A3) we have that

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{j}_{\lambda_n} \notin \mathscr{J}_0\right) = 0.$$

Remark:

- The least we can expect from any reliable model selection procedure is the probability of underfitting tends to zero.
- Under reasonable conditions almost all model selection criteria such as the AIC, Mallow's C_p or the BIC have a nil asymptotic probability of underfitting.

Limiting Overfitting Probability

Theorem 4:

Let $j \in \mathscr{J}$, then under Assumptions (A1), (A2) and (A3) we have that

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{j}_{\lambda_n} = j\right) = \lim_{n\to\infty} \mathbb{1}_{j\in\mathscr{J}_0}\ \alpha^{\lambda_n}_{j-\jmath_0,K-\jmath_0}$$

where $\mathbb{1}$ denotes the indicator function. The quantity $\alpha^{\lambda_n}_{j-\jmath_0,K-\jmath_0}$ can be defined (and computed) as follow:

- Let $Z_l^2 \overset{iid}{\sim} \chi^2_1,\ \ l = 1, ..., K + j + 1.$
- Let $\mathcal{W}_l = Z_l^2 - Z_{l+1}^2,\ l = 1, 2, ..., K + j.$

Then we have that:

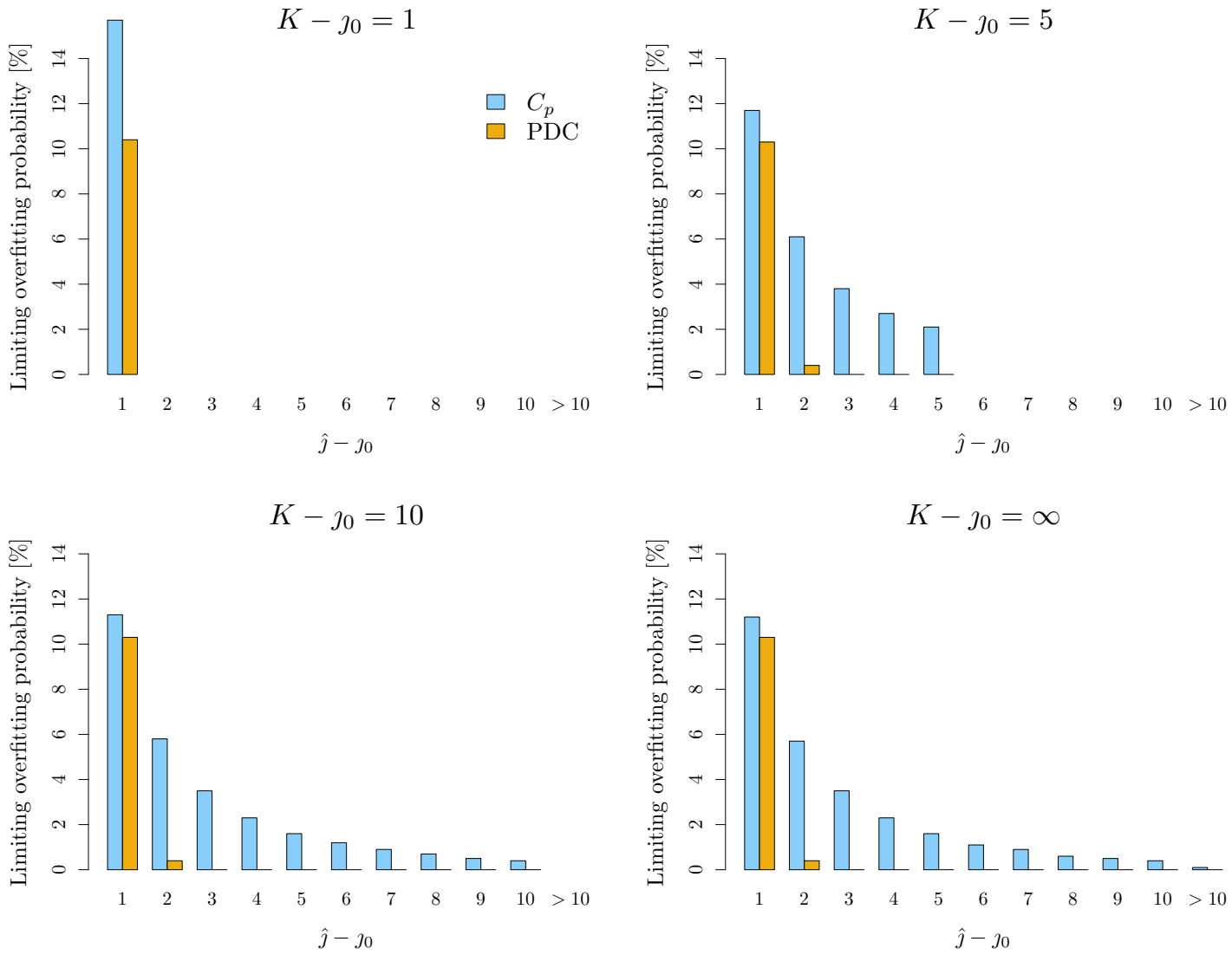
$$\alpha^{\lambda_n}_{j-\jmath_0,K-\jmath_0} = \mathbb{P}\big(\mathcal{W}_1 \leq -\lambda_n(j-\jmath_0),\ ...,\ \mathcal{W}_{K+j} \leq \lambda_n(K-\jmath_0)\big).$$

Limiting Overfitting Probability

Comparing the PDC and the C_p:

- The choice $\lambda_n = 2$ corresponds to the value of λ_n determined by Theorem 1 and, in this situation, **the PDC is the "equivalent" in the PDC class to Mallow's C_p in the classical approach**.
- *Woodroofe, 1982* and later *Zhang, 1992* derived (under slightly different conditions) the **limiting overfitting probability for C_p** (and asymptotically equivalent methods).

Limiting Overfitting Probability



Limiting Overfitting Probability

Remarks:

From Theorem 4 we can show that for $\lambda_n = 2$:

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{\jmath}_2 = \jmath_0\right) \geq \lim_{K-\jmath_0\to\infty} \lim_{n\to\infty} \mathbb{P}\left(\hat{\jmath}_2 = \jmath_0\right) \approx 0.894,$$

$$\lim_{n\to\infty} \mathbb{E}\left[\hat{\jmath}_2\right] \leq \lim_{K-\jmath_0\to\infty} \lim_{n\to\infty} \mathbb{E}\left[\hat{\jmath}_2\right] \approx \jmath_0 + 0.111.$$

For the C_p (and asymptotically equivalent methods) *Woodroofe (1982)* showed that:

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{\jmath}_{C_p} = \jmath_0\right) \geq \lim_{K-\jmath_0\to\infty} \lim_{n\to\infty} \mathbb{P}\left(\hat{\jmath}_{C_p} = \jmath_0\right) \approx 0.712,$$

$$\lim_{n\to\infty} \mathbb{E}\left[\hat{\jmath}_{C_p}\right] \leq \lim_{K-\jmath_0\to\infty} \lim_{n\to\infty} \mathbb{E}\left[\hat{\jmath}_{C_p}\right] \approx \jmath_0 + 0.946.$$

Efficiency and Consistency

Theorem 5:

Under Assumptions (A.1), (A.2) and (A.3) we have that:

$$\frac{L_n\left(\hat{j}_{\lambda_n}\right)}{L_n\left(j_0\right)} \xrightarrow{\mathcal{P}} 1,$$

where

$$L_n(j) = \frac{||\boldsymbol{\mu} - \hat{\boldsymbol{y}}_j||_2^2}{n}$$

Therefore, the PDC selection procedure is **asymptotically loss efficient** (*Shao, 1997*). If in addition $lim_{n\to\infty}\lambda_n = \infty$ then we also have that:

$$\mathbb{P}\left(\hat{j}_{\lambda_n} = j_0\right) \xrightarrow{\mathcal{P}} 1.$$

In this case the procedure is also **consistent**.

Asymptotic Properties: The Unordered Case

Iterative rule:

If covariates are "unordered" we propose applying the following rule:

$$\underset{k=1,\ldots,K-j}{\operatorname{argmax}} \; ||\hat{\boldsymbol{y}}_j - \hat{\boldsymbol{y}}_{j+1}^{(k)}||_2^2 \,.$$

Theorem 6:

Let $\mathbf{X}^\star$ denote the matrix $\mathbf{X}$ whose columns are reorganised according to the above iterative rule. Then under Assumption (A.3) the first j_0 columns of $\mathbf{X}^\star$ contain all significant elements of $\boldsymbol{\beta}$.

Remark:

Other ordering rules can in principle be applied such as e.g. the lasso sequence (see e.g. *Donoho, 2006*).

Simulation Study

Setting:

- We consider linear models: $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma_{\varepsilon}^{2}\mathbf{I}\right)$ and where the matrix $\mathbf{X}$ is randomly generating for each replication such that $\mathbf{x}_i \sim \mathcal{N}\left(\mathbf{0}_p, \boldsymbol{\Sigma}\right)$.
- We compare the "performance" of the following procedures:
  - $\widehat{\text{PDC}}$ (i.e. $\lambda_n = 2$), $\widehat{\text{PDC}}^{\bullet}$ (i.e. $\lambda_n = \log(n)$) and $\widehat{\text{PDC}}^{\star}$ (i.e. $\lambda_n = 2\log(\log(n))$).
  - FPE (*Akaike, 1969*), AIC (*Akaike, 1974*), AICc (*Hurvich and Tsai, 1989*), FPEu and AICu (*McQuarrie et al., 1997*).
  - BIC (*Schwarz (1978)*), HQ (*Hannan and Quinn (1979)*) and HQc (*McQuarrie and Tsai (1998)*).
  - Lasso (*Tibshirani, 1996*) and Elastic Net (*Zou & Hastie, 2005*).

Simulation Study

Models:

- Model 1:

$$\boldsymbol{\beta} = (0, 1, 0, 1, 0, 1, 0, 1, 0, 1, \underbrace{0, ..., 0}_{50})$$

with $\sigma_\epsilon^2 = 1$, $\text{corr}(\mathbf{x}_j, \mathbf{x}_k) = 0.5^{|j-k|}$ and $n = 80$.

- **Model 2**:

$$\boldsymbol{\beta} = (0.3, 0, 0.3, 0, 0.3, 0, 0.3, 0, 0.3, 0)$$

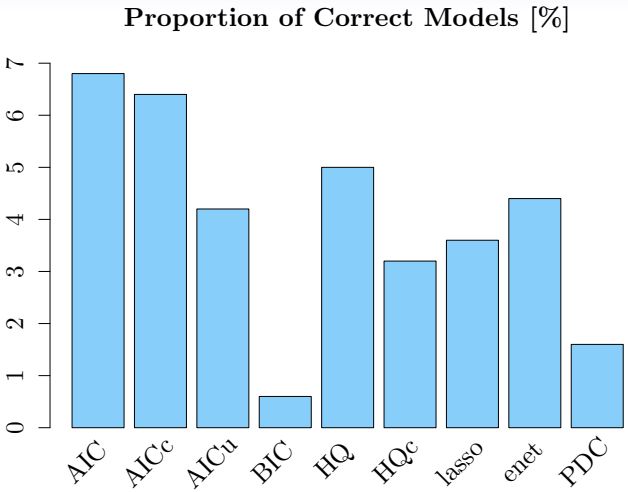
with $\sigma_\epsilon^2 = 1$, $\text{corr}(\mathbf{x}_j, \mathbf{x}_k) = 0.75^{|j-k|}$ and $n = 100$.

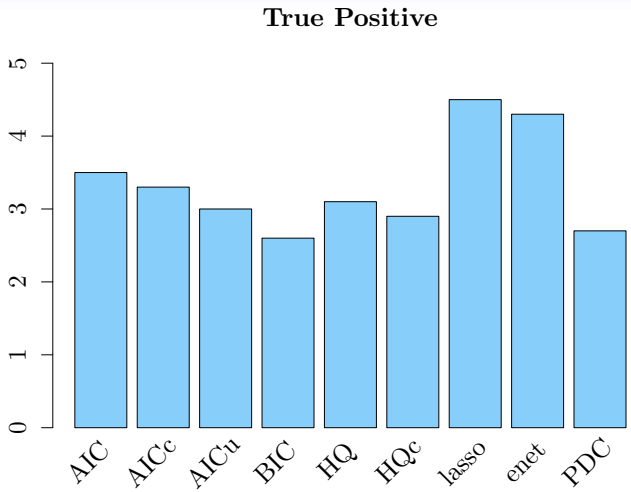
- Model 3:

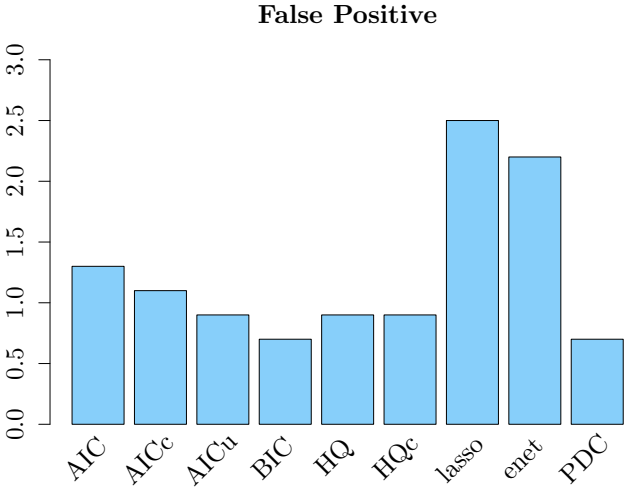
$$\boldsymbol{\beta} = (2, 0, 1, 2, 0, 1, \underbrace{0, ..., 0}_{16}, \underbrace{0.1, ..., 0.1}_{6}, \underbrace{0, ..., 0}_{16}, 2, 0, 1, 2, 0, 1)$$

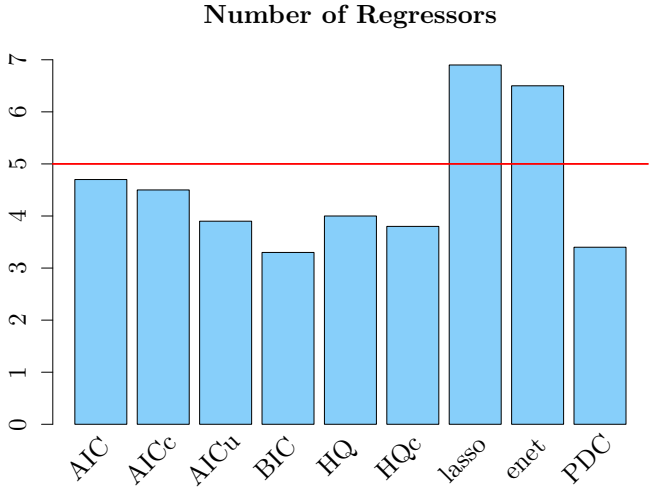
with $\sigma_\epsilon^2 = 4$, $\text{corr}(\mathbf{x}_j, \mathbf{x}_k) = 0.5^{|j-k|}$ and $n = 100$.

Model 2

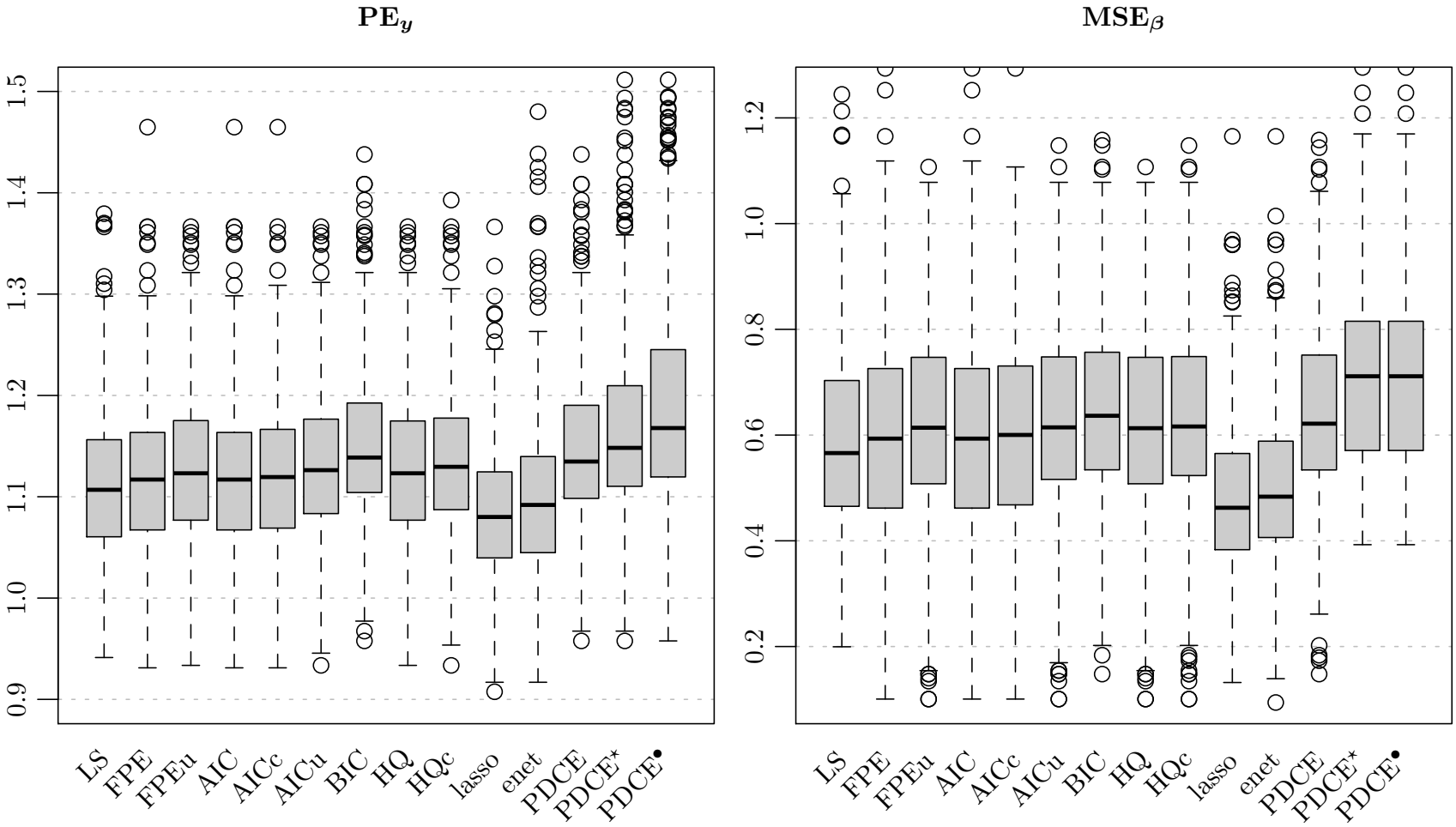

Proportion of Correct Models [%]
0 1 2 3 4 5 6 7
AIC AICc AICu BIC HQ HQc lasso enet PDC


True Positive
0 1 2 3 4 5
AIC AICc AICu BIC HQ HQc lasso enet PDC


False Positive
0.0 0.5 1.0 1.5 2.0 2.5 3.0
AIC AICc AICu BIC HQ HQc lasso enet PDC


Number of Regressors
0 1 2 3 4 5 6 7
AIC AICc AICu BIC HQ HQc lasso enet PDC

Model 2


$\mathbf{PE}_y$
$\mathbf{MSE}_\beta$
LS
FPE
FPEu
AIC
AICc
AICu
BIC
HQ
HQc
lasso
enet
PDCE
PDCE⋆
PDCE•

Simulation Study

Models:

- Model 1:

$$\boldsymbol{\beta} = (0, 1, 0, 1, 0, 1, 0, 1, 0, 1, \underbrace{0, ..., 0}_{50})$$

with $\sigma_\epsilon^2 = 1$, $\text{corr}(\mathbf{x}_j, \mathbf{x}_k) = 0.5^{|j-k|}$ and $n = 80$.

- Model 2:

$$\boldsymbol{\beta} = (0.3, 0, 0.3, 0, 0.3, 0, 0.3, 0, 0.3, 0)$$

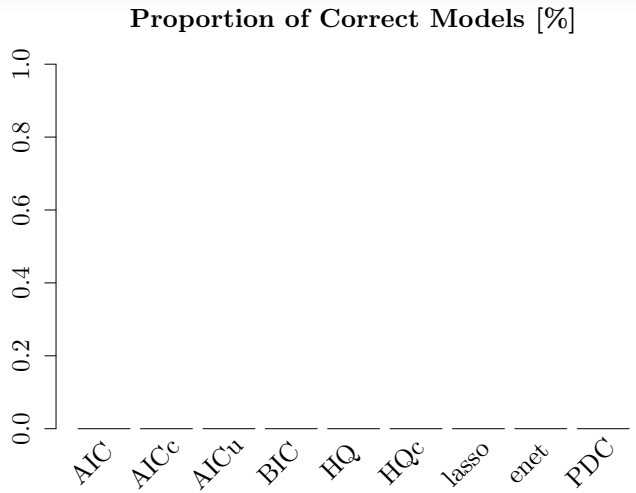
with $\sigma_\epsilon^2 = 1$, $\text{corr}(\mathbf{x}_j, \mathbf{x}_k) = 0.75^{|j-k|}$ and $n = 100$.

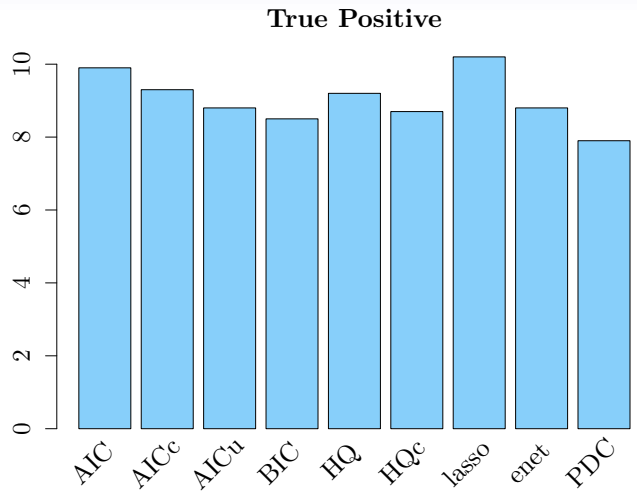
- **Model 3**:

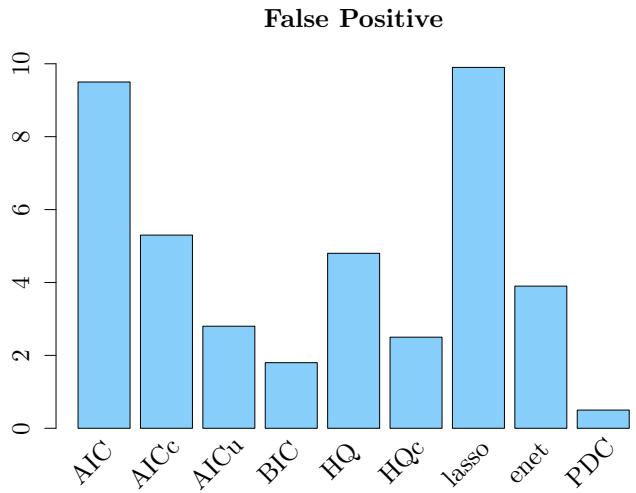
$$\boldsymbol{\beta} = (2, 0, 1, 2, 0, 1, \underbrace{0, ..., 0}_{16}, \underbrace{0.1, ..., 0.1}_{6}, \underbrace{0, ..., 0}_{16}, 2, 0, 1, 2, 0, 1)$$

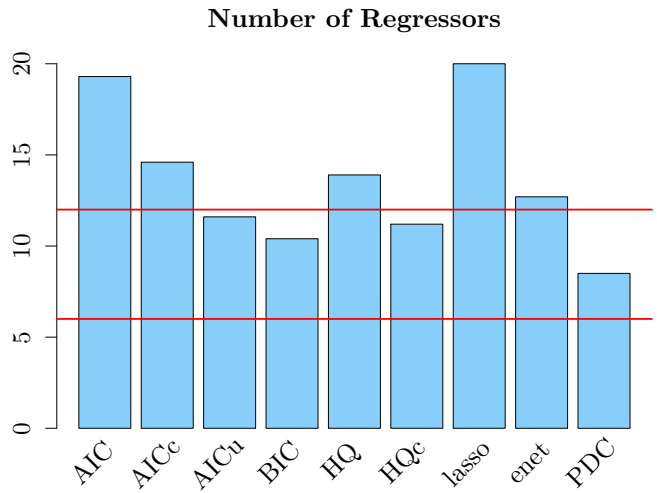
with $\sigma_\epsilon^2 = 4$, $\text{corr}(\mathbf{x}_j, \mathbf{x}_k) = 0.5^{|j-k|}$ and $n = 100$.

Model 3

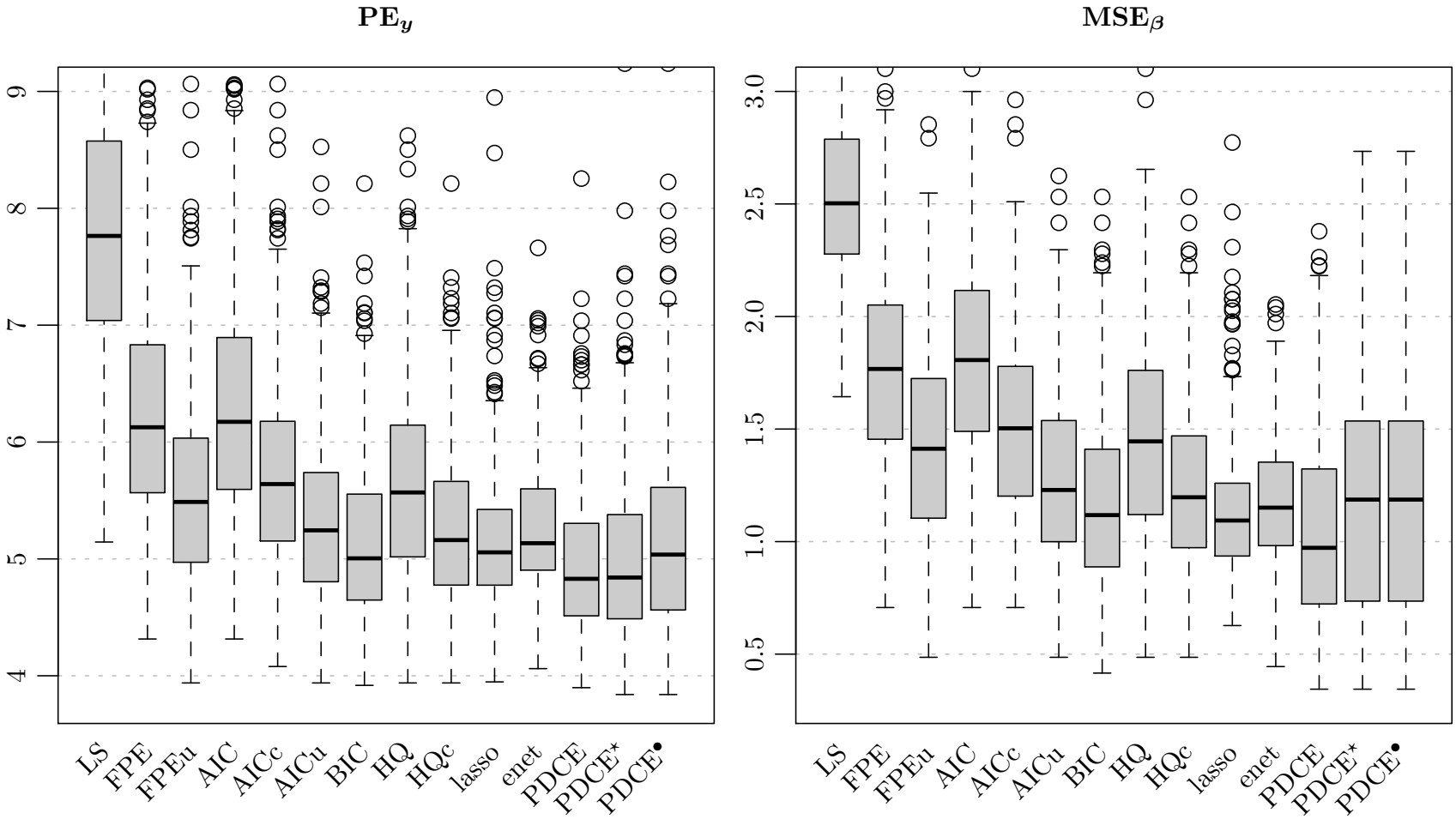

Proportion of Correct Models [%]
0.0
0.2
0.4
0.6
0.8
1.0
AIC
AICc
AICu
BIC
HQ
HQc
lasso
enet
PDC


True Positive
0
2
4
6
8
10
AIC
AICc
AICu
BIC
HQ
HQc
lasso
enet
PDC


False Positive
0
2
4
6
8
10
AIC
AICc
AICu
BIC
HQ
HQc
lasso
enet
PDC


Number of Regressors
0
5
10
15
20
AIC
AICc
AICu
BIC
HQ
HQc
lasso
enet
PDC

Model 3


$\mathbf{PE}_y$
$\mathbf{MSE}_\beta$
LS
FPE
FPEu
AIC
AICc
AICu
BIC
HQ
HQc
lasso
enet
PDCE
PDCE*
PDCE•

Other Applications and Extensions of the PDC Approach

Extensions:

- Generalized Linear Model
- Smoothing Splines:
  - Choosing between a linear and a nonlinear model.
- Order selection in AR models.
- Random effect selection in Mixed Linear Models:
  - The "complexity" of such models is not a "well defined" quantity.
  - PDC appears to outperform conditional AIC (cAIC) of *Vaida & Blanchard (2005)*.
- Large and High-Dimensional Problems ($n \ll p$):
  - Computational "short-cuts".
  - Additional assumptions are required.

Logistic Regression

Setting:

Logistic regression fits a model of the form:

$$\pi_i = \frac{1}{1 + \exp\left(-\mathbf{x}_i^T \boldsymbol{\beta}\right)}, \quad i = 1, ..., n$$

to an observed vector of binary data $\mathbf{y}$. Given an estimator of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$ we can the prediction $\hat{\eta}_i$ as:

$$\hat{\eta}_i = \begin{cases} 1 & \text{if } \hat{\pi}_i > C_0 \\ 0 & \text{if } \hat{\pi}_i \leq C_0 \end{cases}$$

for some cutoff C_0 (typically 0.5).

Logistic Regression

PDC for logistic regression:

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two nested models associated to the prediction vectors $\hat{\boldsymbol{\eta}}_{(1)}$ and $\hat{\boldsymbol{\eta}}_{(2)}$. We consider the following Prediction Divergence Criterion:

$$\mathrm{C} = \mathbb{E}\left[\mathbb{E}_0\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\hat{\eta}^{\mathbf{0}}_{(\mathbf{1})i}\neq\hat{\eta}_{(\mathbf{2})i}}\right]\right].$$

Using Theorem 1 we may construct an estimator of C as:

$$\hat{\mathrm{C}} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\hat{\eta}_{(\mathbf{1})i}\neq\hat{\eta}_{(\mathbf{2})i}} + \hat{\omega}_{1,2}$$

where $\hat{\omega}_{1,2}$ is the estimated optimism which are conveniently approximated using the results of *Efron, 1986*.
The covariates are "ordered" according to the following rule:

$$\underset{k=1,\ldots,K-j}{\operatorname{argmax}} \; ||\hat{\boldsymbol{\pi}}_{(j)} - \hat{\boldsymbol{\pi}}_{(j+1)^{(k)}}||_2^2 .$$

Acute Leukemia Class Prediction

- **Objective of class prediction using DNA microarrays**: identify the smallest possible set of genes that can still achieve good predictive performance.
- **Example: Acute leukemia**. Variability in clinical outcome and treatments. Subtypes:
  - acute lymphoblastic leukemia (ALL)
  - acute myeloid leukemia (AML);
- Cancer classification is central to cancer treatment.
- We consider the data of *Golub et al., 1999*.
- The leukaemia data consist of **7129 genes** and 72 samples.
- In the training data set, there are 38 samples, among which 27 ALL and 11 AML. The remaining 34 samples are used to test the prediction accuracy of the diagnostic rule.

Acute Leukemia Class Prediction

We compare the performance of the following procedures:

- PDC for logistic regression (0/1 loss function).
- Golub's original method (see *Golub et al., 1999*).
- Support vector machine + recursive feature elimination (see *Guyon et al., 2002*).
- Penalised logistic regression + recursive feature elimination (see *Zhu & Hastie, 2004*).
- Nearest shrunken centroids (see *(Tibshirani et al., 2002*).
- Elastic net (see *Zou & Hastie, 2005*).

Acute Leukemia Class Prediction

Results:

Table: The results are taken from *Zou & Hastie, 2005*, except for the PDC.

Method	Tenfold CV error	Test error	Number of genes
Golub	3/38	4/34	50
Support vector machine*	2/38	1/34	31
Penalised logistic regression*	2/38	1/34	26
Nearest shrunken centroids	2/38	2/34	21
Elastic net	3/38	0/34	45
PDC (GLM + 0/1 loss)	1/38	1/34	6

Summary

1. New class of error measures:
   - flexible framework for deriving model selection criteria,
   - examples: PDC, GMM, robust statistics, ...
2. New class of model selection criteria:
   - less prone to overfitting than "classical" counterpart,
   - asymptotically loss efficient and possibly consistent,
   - performs particularly well in "sparse" settings.
3. Computationally efficient.
4. Future $\rightarrow$ generalized linear (mixed) models.

Thank you very much for your attention