ITALIAN STATISTICAL SOCIETY
"SMART STATISTICS FOR SMART APPLICATIONS"

Milan, June 19, 2019

A new tuning parameter selector in lasso regression

Gianluca SOTTILE[1] and Vito MR MUGGEO[1]

[1]Departments of Economics, Business and Statistics
University of Palermo - Italy
gianluca.sottile@unipa.it

Contents

Contents

Background

Let consider a classical regression framework, where n is the number of observations and p the number of covariates, in the low-dimensional ($n > p$) and largely in the high-dimensional context ($n \ll p$), only a small number of variables are truly informative.

Background

Let consider a classical regression framework, where n is the number of observations and p the number of covariates, in the low-dimensional ($n > p$) and largely in the high-dimensional context ($n \ll p$), only a small number of variables are truly informative.

Penalised regression methods

These methods operate maximising the penalised likelihood function

$$\frac{1}{n}\ell(\boldsymbol{\beta}) - P_\lambda(\boldsymbol{\beta}) \tag{1}$$

with respect to $\boldsymbol{\beta} \in \mathbb{R}^p$ and $P_\lambda(\cdot)$ is a penalty function.

Least Absolute Shrinkage and Selection Operator

LASSO (Tibshirani, 1996)

The penalty function reduces to

$$P_\lambda(\cdot) = \lambda \sum_{j=1}^{p} |\beta_j|$$

where $\lambda \geq 0$ is known as tuning parameter

Least Absolute Shrinkage and Selection Operator

LASSO (Tibshirani, 1996)

The penalty function reduces to

$$P_\lambda(\cdot) = \lambda \sum_{j=1}^{p} |\beta_j|$$

where $\lambda \geq 0$ is known as tuning parameter

- allows to perform variable selection
- λ balances the trade-off between model fit and model sparsity
- $\lambda \to 0 \Rightarrow \hat{\boldsymbol{\beta}}_\lambda \to \hat{\boldsymbol{\beta}}^{\text{OLS}}$
- $\lambda \to +\infty \Rightarrow \hat{\boldsymbol{\beta}}_\lambda \to \mathbf{0}$

Least Absolute Shrinkage and Selection Operator

LASSO (Tibshirani, 1996)

The penalty function reduces to

$$P_\lambda(\cdot) = \lambda \sum_{j=1}^{p} |\beta_j|$$

wh

"Which is the optimal tuning parameter?"

- allows to perform variable selection
- λ balances the trade-off between model fit and model sparsity
- $\lambda \to 0 \Rightarrow \hat{\boldsymbol{\beta}}_\lambda \to \hat{\boldsymbol{\beta}}^{\text{OLS}}$
- $\lambda \to +\infty \Rightarrow \hat{\boldsymbol{\beta}}_\lambda \to \mathbf{0}$

Contents

Penalised linear regression framework

Let consider $y \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathrm{I})$ and $\boldsymbol{X}$ the model matrix $(n \times p)$,

- $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ is the unknown mean vector;
- σ^2 is the variance of the error.

Penalised linear regression framework

Let consider $y \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathrm{I})$ and $\boldsymbol{X}$ the model matrix ($n \times p$),

- $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ is the unknown mean vector;
- σ^2 is the variance of the error.

The mean vector is estimated by $\hat{\boldsymbol{\mu}}_\lambda = \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda$, where $\hat{\boldsymbol{\beta}}_\lambda$ is the estimator that minimize the penalised least squares function

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}|\beta_j|. \quad (2)$$

Penalised linear regression framework

Recap of theoretical assumptions of the LASSO framework

Being β^0 the true vector of coefficients and $S_0 := \{j : \beta_j^0 \neq 0\}$ the active set; any selection criteria deliver an estimator $\hat{S}$ of S_0.
We assume:

- the "*beta-min* condition", i.e., $\min_{j \in S_0} |\beta_j^0| \geq c \cdot \sqrt{2\phi \log p}$;
- the true number of nonzero coefficients have to obey to $d_0 \leq n/(2 \log p)$;
- the "irrepresentable condition" or "restricted eigenvalue condition" that is a condition on the model matrix, is a sufficient and necessary condition for consistent variable selection.

Penalised linear regression framework

Recap of previous proposals

$$\text{AIC} = \log(\hat{\sigma}^2_\lambda) + 2\ \mathsf{d}_\lambda n^{-1}$$

$$\text{BIC} = \log(\hat{\sigma}^2_\lambda) + \log n\ \mathsf{d}_\lambda n^{-1}$$

$$\text{EBIC} = \log(\hat{\sigma}^2_\lambda) + (\log n + 2\gamma \log p)\ \mathsf{d}_\lambda n^{-1}$$

$$\text{GCV} = \hat{\sigma}^2_\lambda / \left(1 - \mathsf{d}_\lambda n^{-1}\right)^2$$

$$\text{GIC} = \log(\hat{\sigma}^2_\lambda) + \mathsf{c}_n \log p\ \mathsf{d}_\lambda n^{-1}$$

$$\text{k-fold CV} = \sum_{s=1}^{k} \sum_{(y_s, \boldsymbol{x}_s) \in T^{-s}} \left(y_s - \boldsymbol{x}_s^T \hat{\boldsymbol{\beta}}^{(s)}_\lambda \right)^2$$

Penalised linear regression framework

Recap of previous proposals

$$\mathsf{AIC} = \log(\hat{\sigma}^2_\lambda) + 2\,\mathsf{d}_\lambda n^{-1}$$

where:

- $\hat{\sigma}^2_\lambda = \mathsf{RSS}_\lambda/(n - \mathsf{d}_\lambda)$
- $\gamma > 0$
- c_n is a parameter which depends on n.

$$\text{k-fold CV} = \sum_{s=1}^{k} \sum_{(y_s, \boldsymbol{x}_s) \in T^{-s}} \left(y_s - \boldsymbol{x}_s^T \hat{\boldsymbol{\beta}}_\lambda^{(s)} \right)^2$$

Contents

The proposed criterion

Weighted signal-to-noise ratio (WSNR)

We suggest to select λ as the maximizer of

$$\arg\max_{\lambda} w_{\lambda} \frac{\| \hat{\boldsymbol{\beta}}_{\lambda} \|_1}{\hat{\sigma}_{\lambda}}, \qquad (3)$$

where

- $w_{\lambda} = \mathrm{d}_{\lambda}^{-1}$ is the model degrees of freedom, or the cardinality of the active set, i.e., $|S_{\lambda}| = |\{j : \hat{\beta}_{\lambda,j} \neq 0\}|$;
- σ_{λ} is the square root of the dispersion parameter that could be fixed or estimated.

Contents

Prostate Cancer data set (Stamey et al., 1989)

It is a study on prostate cancer, measuring the correlation between the level of prostate-specific antigen ($y =$ lpsa, log-psa) and:

- $x_1 =$ lcavol (log-cancer volume)
- $x_2 =$ lweight (log-prostate weight)
- $x_3 =$ age (age of patient)
- $x_4 =$ lbhp (log-amount of benign hyperplasia)
- $x_5 =$ svi (seminal vesicle invasion)
- $x_6 =$ lcp (log-capsular penetration)
- $x_7 =$ gleason (Gleason Score)
- $x_8 =$ pgg45 (percent of Gleason scores 4 or 5)

Prostate Cancer data set (Stamey et al., 1989)

Performance assessment

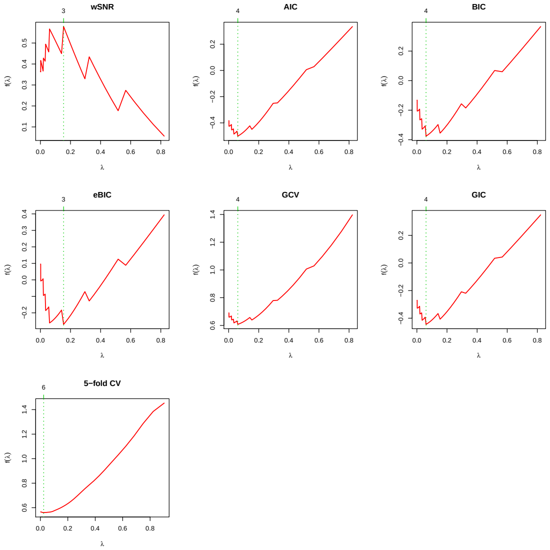
- Training set (ts): $n = 73\ (75\%)$
- Validation set (vs): $n = 24\ (25\%)$
- Number of nonzero coefficients selected
- Prediction error: $\mathsf{PE} = \sum_i (y_{i,\mathsf{vs}} - \boldsymbol{x}_{i,\mathsf{vs}}^T \hat{\boldsymbol{\beta}}_{\mathsf{ts}})^2 / n_{\mathsf{vs}}$

Prostate Cancer data set (Stamey et al., 1989)

Table 1: Tuning parameter selection of the Prostate Cancer data set. The number of nonzero coefficients, the tuning parameter selected (λ^*) and the prediction error are reported.

	Coeff	λ^*	PE
WSNR	3	0.1541	0.3994
AIC	4	0.0608	0.3990
BIC	4	0.0608	0.3990
EBIC	3	0.1541	0.3994
GCV	4	0.0608	0.3990
GIC	4	0.0608	0.3990
CV	6	0.0289	0.4027

Prostate Cancer data set (Stamey et al., 1989)



Diabetes data set (Efron et al., 2003)

It is a study on diabetes. A quantitative measure of disease progression one year after baseline as well as Ten baseline variables are collected by $n = 442$ diabetes patients.

- x_1 = age
- x_2 = sex
- x_3 = body mass index (bmi)
- x_4 = average blood pressure (map)
- $x_{5:10}$ = blood serum measurements (tc, ldl, hdl, tch, ltg, glu)
- $x_{11:64}$ = interaction terms

Diabetes data set (Efron et al., 2003)

Performance assessment

- Training set (ts): $n = 332\ (75\%)$
- Validation set (vs): $n = 110\ (25\%)$
- Number of nonzero coefficients selected
- Prediction error: $\textsf{PE} = \sum_i (y_{i,\textsf{vs}} - \boldsymbol{x}_{i,\textsf{vs}}^T \hat{\boldsymbol{\beta}}_{\textsf{ts}})^2 / n_{\textsf{vs}}$

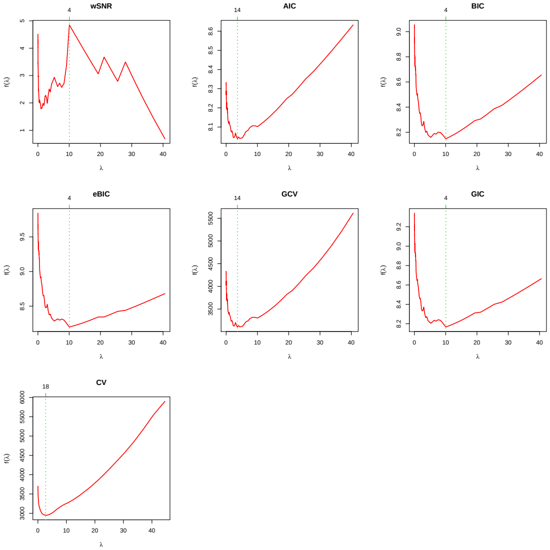
Diabetes data set (Efron et al., 2003)

Pe

Table 2: Tuning parameter selection of the Diabetes data set. The number of nonzero coefficients, the tuning parameter selected (λ^*) and the prediction error are reported.

	Coeff	λ^*	PE
WSNR	4	10.05	3304.13
AIC	14	3.61	3029.81
BIC	4	10.05	3304.13
EBIC	4	10.05	3304.13
GCV	14	3.61	3029.81
GIC	4	10.05	3304.13
CV	18	2.73	3042.47

Diabetes data set (Efron et al., 2003)



Pe

Contents

To sum-up

- We proposed a new criterion to choose the regularization parameter;

To sum-up

- We proposed a new criterion to choose the regularization parameter;
- Our proposal can be extended to generalized linear models, e.g., Poisson and logistic regression;

To sum-up

- We proposed a new criterion to choose the regularization parameter;
- Our proposal can be extended to generalized linear models, e.g., Poisson and logistic regression;
- We applied our proposal to prostate cancer data and our proposal was able to select, three non-zero covariates log-cancer volume, log-cancer weight and seminal vesicle invasion;

To sum-up

- We proposed a new criterion to choose the regularization parameter;
- Our proposal can be extended to generalized linear models, e.g., Poisson and logistic regression;
- We applied our proposal to prostate cancer data and our proposal was able to select, three non-zero covariates log-cancer volume, log-cancer weight and seminal vesicle invasion;
- We applied our proposal to diabetes data and our proposal was able to select, four non-zero covariates body mass index, average blood pressure, hdl and ltg (blood serum measurements).

Thanks for the attention!!!