Sparse Conformal Predictor

Mohamed Hebiri
hebiri@math.jussieu.fr

LPMA
Paris 7-Diderot

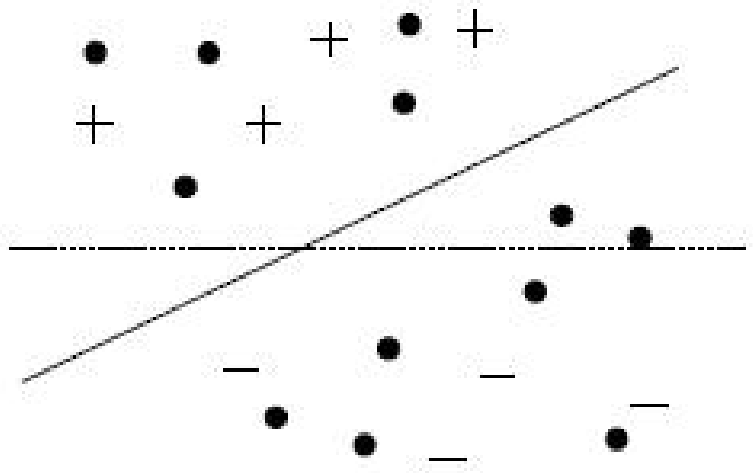
Rennes - 29 Aout 2008

Outline

Transductive setting

References:

- Vapnik ’98
- Joachims ’99



Linear Regression Model

Observations: $\mathcal{E}_n = \{(x_1, y_1), \ldots, (x_n, y_n), x_{new}\}$

$$y_i = x_i \beta^* + \xi_i$$

- Covariates vector: $x_i = (x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^p, \quad i \geq 1$
- New observation: $x_{new} \in \mathbb{R}^p$
- Response: $y_i \in \mathbb{R}, \quad i \geq 1$
- Unknown parameter: $\beta^* = (\beta_1^*, \ldots, \beta_p^*)' \in \mathbb{R}^p$
- Noise: $\xi_i \sim \mathcal{N}(0, \sigma^2)$

Objectives

Goal I: Given $\mathcal{E}_n$ and $\varepsilon > 0$, provide a *confidence predictor* (interval) Γ^ε in which y_{new} lies with probability $1 - \varepsilon$

Tool: Conformity measure between x_{new} and the previous x_i's

- distance (geometric, neighborhood,...)
- *distance of similarity*: to be defined later !

Goal II: Exploit the *sparsity* of the model (β^* contains many zero components) if necessary

Tool: Use a variable selection procedure (LASSO,...)

Conformal prediction: Vovk et al. '05

Notation:

- $y \in \mathbb{R}$: potential value of y_{new}
- $|\mathcal{A}|$: cardinal of the set $\mathcal{A}$

Nonconformity score $\alpha(y) = (\alpha_1(y), \ldots, \alpha_n(y), \alpha_{new}(y))'$

- $\alpha_i(y)$: similirity between (x_{new}, y) and (x_i, y_i)
- relative information: *p-value*

$$p(y) = \frac{1}{n+1} \left| \{ i \in \{1, \ldots, n, new\} : \alpha_i(y) \leq \alpha_{new}(y) \} \right|$$

- $p(y) \in [\frac{1}{n+1}; 1]$
- the smaller $p(y)$ is, the less likely the tested pair (x_{new}, y) is (this choice reduces y to be an outlier)

Conformal predictor Γ^{ε}: labels $y \in \mathbb{R}$ such that $p(y) > \varepsilon$.

LASSO estimator: Tibshirani '96

LASSO estimator

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{Argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

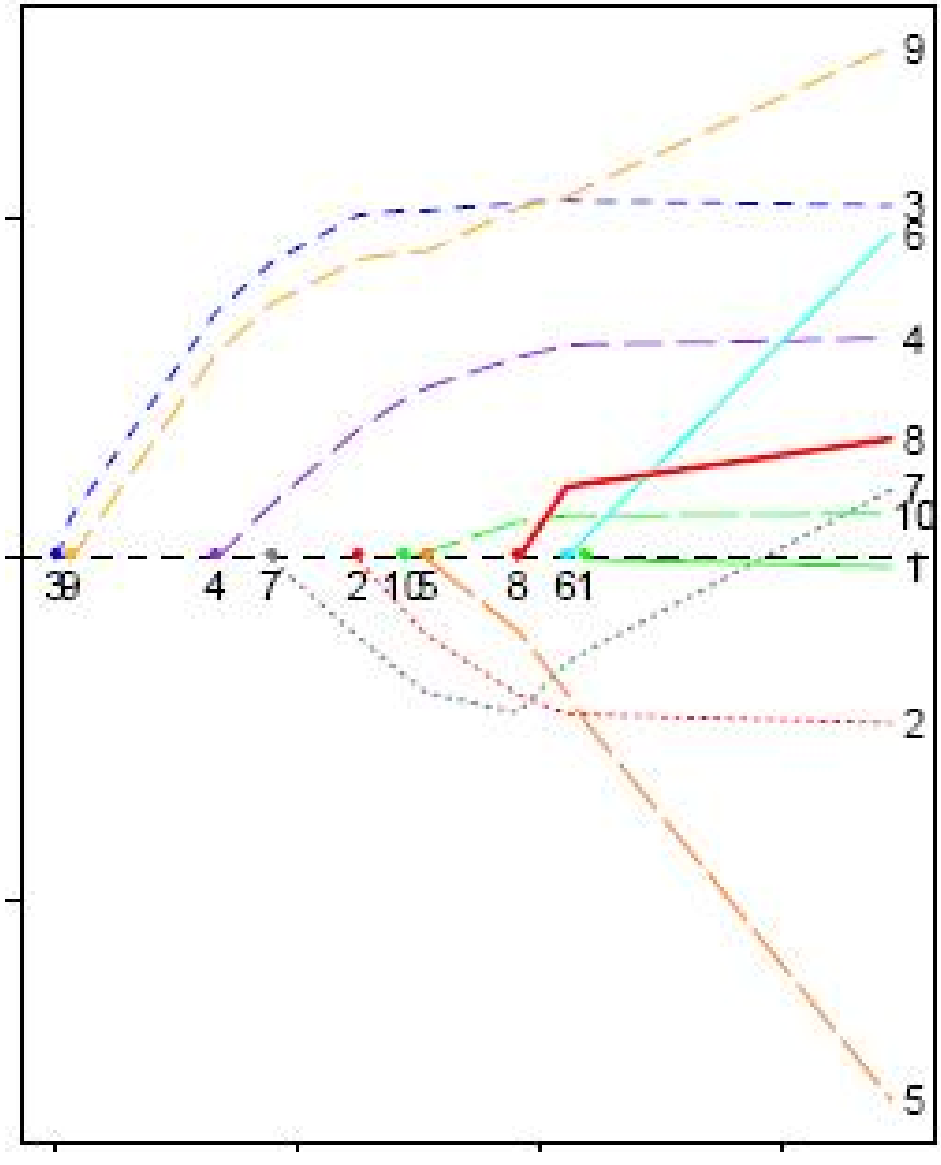
- Tuning parameter: λ

Motivation:

- Sparse solution $\hat{\beta}$ (i.e., many coefficients reduced to 0)
- Interpretable results whenever the model is sparse

Solution approximation: **LARS** algorithm (Efron et al. '04)

- Example: available Diabetes dataset (10 covariates)
- $\hat{\beta}_{\lambda_1}, \ldots, \hat{\beta}_{\lambda_K}$: approximations of the LASSO solution at the transition points $\lambda = \lambda_1, \ldots, \lambda_K$



- Step k: $\hat{\mu}_k = \mathbf{x}_k \hat{\beta}_{\lambda_k} = \mathbf{x}_k (\mathbf{x}_k' \mathbf{x}_k)^{-1} (\mathbf{x}_k' \mathbf{y} - \frac{\lambda_k}{2} s_k)$
  - response vector: $\mathbf{y} = (y_1, \ldots, y_n)'$
  - design matrix: $\mathbf{x} = (x_1', \ldots, x_n')'$
  - sign vector: s_k
  - $\mathbf{x}_k$ is the restriction of $\mathbf{x}$ to the columns corresponding to selected covariates

Does not take into account x_{new} while constructing $\hat{\beta}$!

Sparse Conformal Predictor

- Consider the augmented dataset: $\widetilde{\mathbf{x}} = (x_1', \ldots, x_n', x_{new}')'$ and $\widetilde{\mathbf{y}} = (y_1, \ldots, y_n, y)'$
- For **each** transition point λ_k, define the LASSO estimator $\hat{\mu}_k$ based on $\widetilde{\mathbf{x}}_k$ and $\widetilde{\mathbf{y}}$

Define the **Nonconformity score**

$$\alpha^k(y) \quad := \quad |\widetilde{\mathbf{y}} - \hat{\mu}_k| = |A_k + C_k + B_k\, y|$$

where $|\cdot|$ is meant here componentwise and

$$\begin{cases} A_k = (a_1^k, \ldots, a_n^k, a_{new}^k)' := (\mathbf{I} - \mathbf{H}_k)\,(y_1, \ldots, y_n, 0)' \\ B_k = (b_1^k, \ldots, b_n^k, b_{new}^k)' := (\mathbf{I} - \mathbf{H}_k)\,(0, \ldots, 0, 1)' \\ C_k = (c_1^k, \ldots, c_n^k, c_{new}^k)' := \frac{\lambda_k}{2} \widetilde{\mathbf{x}}_k \left(\widetilde{\mathbf{x}}_k' \widetilde{\mathbf{x}}_k\right)^{-1} s_k \end{cases}$$

- The $\alpha^k(y)$ are piecewise linear

Sparse Conformal Predictor

- *p-value:* $p^k(y) = \frac{1}{n+1} \left| \left\{ i : \alpha_i^k(y) \leq \alpha_{new}^k(y) \right\} \right|$
- Predictor at step k: $\Gamma_k^\varepsilon = \{y \in \mathbb{R} : \quad p^k(y) > \varepsilon\}$.

Proposition

Points y such that $\alpha_i^k(y) = \alpha_{new}^k(y)$ *exist*
i) if $b_i^k \neq b_{new}^k$*: when y equals*

$$-\frac{a_i^k - a_{new}^k + c_i^k - c_{new}^k}{b_i^k - b_{new}^k} \quad \text{and} \quad -\frac{a_i^k + a_{new}^k + c_i^k + c_{new}^k}{b_i^k + b_{new}^k}.$$

ii) if $b_i^k = b_{new}^k \neq 0$*: when y equals*

$$-\frac{a_i^k + a_{new}^k + c_i^k + c_{new}^k}{2b_i^k}$$

Conformal Lasso Predictor Γ_ε^F: the smallest Γ_ε^k

Extension

Estimator of the form:

$$\hat{\mu} = u(\widetilde{\mathbf{x}}, s)\widetilde{\mathbf{y}} + v(\widetilde{\mathbf{x}}, s)$$

where $u(\cdot)$ and $v(\cdot)$ are piecewise constant functions w.r.t. $\widetilde{\mathbf{y}}$

We are interested in

- CoLP: $u(\widetilde{\mathbf{x}}, s) = \widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k)^{-1}\widetilde{\mathbf{x}}_k'$
  $v(\widetilde{\mathbf{x}}, s) = -\lambda_k\widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k)^{-1}s_k$
- CoRP: $u(\widetilde{\mathbf{x}}, s) = \widetilde{\mathbf{x}}(\widetilde{\mathbf{x}}'\widetilde{\mathbf{x}} + \mu\mathbf{I}_p)^{-1}\widetilde{\mathbf{x}}'$ and $v = 0$
- CENeP: $u(\widetilde{\mathbf{x}}, s) = \widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k + \mu_k\mathbf{I}_k)^{-1}\widetilde{\mathbf{x}}_k'$
  $v(\widetilde{\mathbf{x}}, s) = -\lambda_k\widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k)^{-1}s_k$

Experimental setting

- All simulations are based on $M = 1000$ replications
- Performance measures:
  - Accuracy: length of intervals
  - Validity criterion: $\text{VAL}^{\varepsilon} = M^{-1} \sum_{m=1}^{M} \mathbb{I}(y_{new} \in (\Gamma_{\varepsilon}^{F})_m)$
  - Variable selection: recovery of the support of β^*

Benchmarks:

- Variable selection: Original LASSO (Tibshirani '96) and the Original Elastic-Net (Zou & Hastie '05) (based on the BIC criterion)
- Accuracy and validity: CoRP (Vovk et al. '05)

Data $p = 50$

$\mathcal{A}^* = \{j : \beta_j^* \neq 0\}$: set of relevant covariates

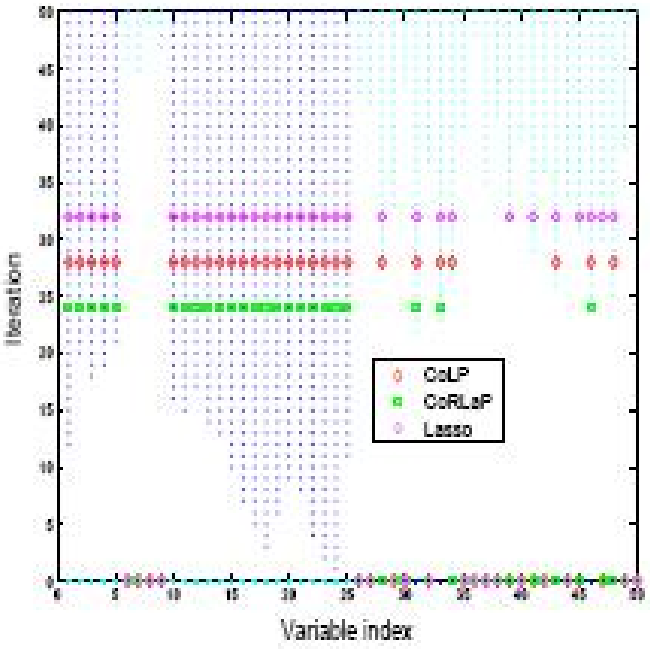
- Example(a): $\mathcal{A}^* = \{1\}$; exponentially decreasing correlations between successives covariates $\{15, \ldots, 35\}$
- Example(b): $\mathcal{A}^* = \{1, \ldots, 5\} \cup \{10, \ldots, 25\}$; correlations as in Example(a)
- Example(c): $\mathcal{A}^* = \{1, \ldots, 15\}$; three groups of highly correlated variables: $G_1 = \{1, \ldots, 5\}$, $G_2 = \{6, \ldots, 10\}$ and $G_1 = \{11, \ldots, 15\}$
- Example(d): $\mathcal{A}^* = \{1, \ldots, p\}$; exponentially decreasing correlations between successives covariates $\{1, \ldots, p\}$

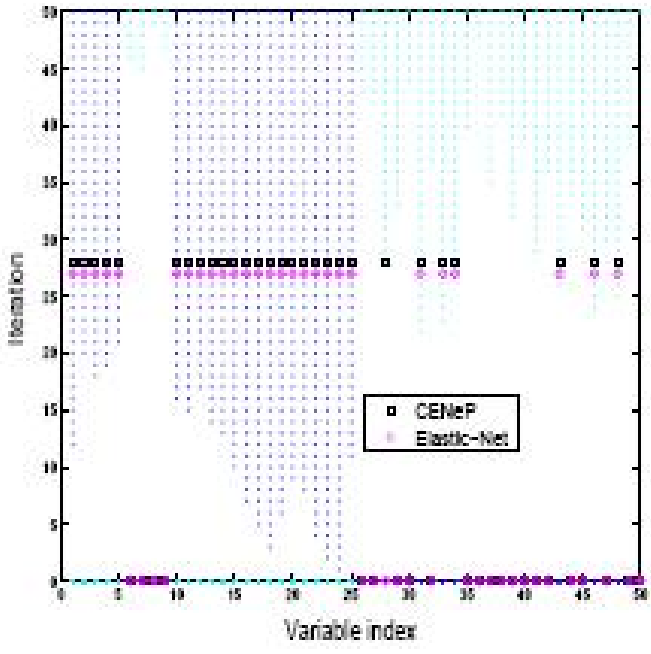
Validity

Table: VAL$^{\varepsilon}$ evaluations

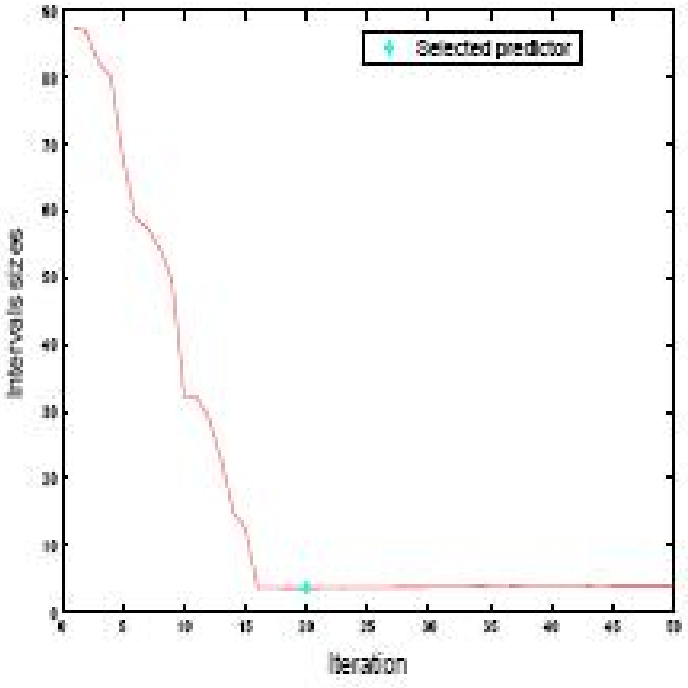
EXAMPLE[n/σ]	CORP	COLP	COLARP	CENEP
EX (A)[300/1]	0.90± 0.02	0.88± 0.02	0.85± 0.02	0.88± 0.02
EX (A)[300/7]	0.89± 0.02	0.91± 0.02	0.89± 0.02	0.90± 0.02
EX (A)[300/15]	0.89± 0.02	0.89 ± 0.02	0.88± 0.02	0.89± 0.02
EX (B)[300/1]	0.90± 0.02	0.88± 0.02	0.87± 0.02	0.87± 0.02
EX (C)[300/1]	0.90± 0.02	0.90± 0.02	0.89± 0.02	0.90± 0.02
EX (D)[300/1]	0.89± 0.02	0.90± 0.02	0.90± 0.02	0.90± 0.02
EX (A)[50/3]	0.89± 0.02	0.67± 0.03	0.41± 0.03	0.79± 0.02
EX (A)[20/3]	0.86± 0.02	0.60± 0.03	0.30± 0.03	0.69± 0.03
EXAMPLE[n/σ]	CORP	COLP	STOPPED-COLP	2-PN-COLP
EX (A)[50/7]	0.85± 0.02	0.62± 0.03	0.82± 0.02	0.88± 0.02
EX (B)[50/1]	0.88± 0.02	0.56± 0.03	0.82± 0.02	0.91 ± 0.02
EX (C)[20/15]	0.88± 0.02	0.61± 0.03	0.77± 0.03	0.90± 0.02
EX (D)[20/1]	0.90± 0.02	0.60± 0.03	0.79± 0.02	0.89± 0.02

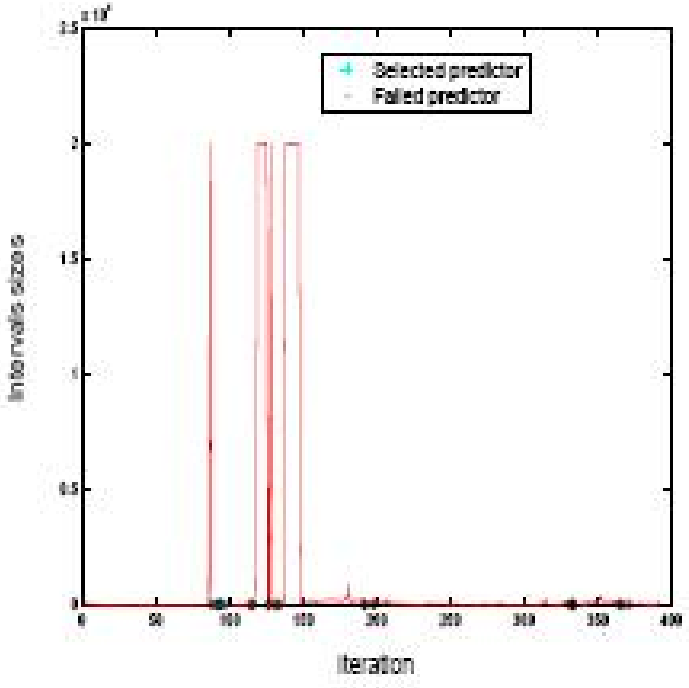
Variable selection: Example(b)[300/5]





Accuracy: Example(b)[$n/5$]





Conclusion

- Theoretical validity (Vovk et al. ’05)
- Theoretical variable selection consistency when selection is based on the accuracy criterion.