High-dimensional regression with unknown variance

Christophe Giraud

Ecole Polytechnique

march 2012

Setting

Gaussian regression with unknown variance:

- $Y_i = f_i + \varepsilon_i$ with $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$
- $f = (f_1, \ldots, f_n)^*$ and σ^2 are unknown
- we want to estimate f

Ex 1 : sparse linear regression

- $f = X\beta$ with β "sparse" in some sense and $X \in \mathbb{R}^{n \times p}$ with possibly $p > n$

Ex 2 : non-parametric regression

- $f_i = F(x_i)$ with $F : \mathcal{X} \to \mathbb{R}$

A plethora of estimators

Sparse linear regression

- **Coordinate sparsity:** Lasso, Dantzig, Elastic-Net, Exponential-Weighting, Projection on subspaces $\{V_\lambda : \lambda \in \Lambda\}$ given by PCA, Random Forest, etc.
- **Structured sparsity:** Group-lasso, Fused-Lasso, Bayesian estimators, etc

Non-parametric regression

- Spline smoothing, Nadaraya kernel smoothing, kernel ridge estimators, nearest neighbors, L^2-basis projection, Sparse Additive Models, etc

Important practical issues

Which estimator should be used?

- **Sparse regression :** Lasso? Random-Forest? Exponential-Weighting?
- **Non-parametric regression :** Kernel regression? (which kernel?) Spline smoothing?

Which "tuning" parameter?

- which penalty level for the lasso?
- which bandwith for kernel regression?
- etc

The objective

Difficulties

- No procedure is universally better than the others
- A sensible choice of the tuning parameters depends on
  - some unknown characteristics of f (sparsity, smoothness, etc)
  - the unknown variance σ^2.

Ideal objective

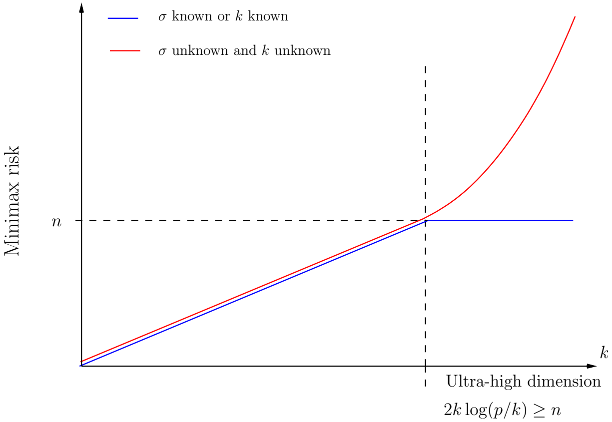
- Select the "best" estimator among a collection $\{\hat{f}_\lambda, \lambda \in \Lambda\}$.

(alternative objective: combine at best the estimators)

Impact of not knowing the variance

Impact of the unknown variance?

Case of coordinate-sparse linear regression



Minimax prediction risk over k-sparse signal as a function of k

Ultra-high dimensional phenomenon

Theorem (N. Verzelen EJS 2012)

When σ^2 is unknown, there exist designs $\mathbf{X}$ of size $n \times p$ such that for any estimator $\widehat{\beta}$, we have either

$$\sup_{\sigma^2>0} \mathbb{E}\left[\|\mathbf{X}(\widehat{\beta}-0_p)\|^2\right] > C_1 n\sigma^2 \ , \qquad \text{or}$$

$$\sup_{\substack{\beta_0 \ k\text{-sparse} \\ \sigma^2>0}} \mathbb{E}\left[\|\mathbf{X}(\widehat{\beta}-\beta_0)\|^2\right] > C_2 k \log\left(\frac{p}{k}\right) \exp\left[C_3 \frac{k}{n} \log\left(\frac{p}{k}\right)\right] \sigma^2.$$

Consequence

When σ^2 unknown, the best we can expect to have is

$$\mathbb{E}\left[\|\mathbf{X}(\widehat{\beta}-\beta_0)\|^2\right] \leq C \inf_{\beta \neq 0} \left\{\|\mathbf{X}(\beta-\beta_0)\|_2^2 + \|\beta\|_0 \log(p)\sigma^2\right\}$$

for any $\sigma^2 > 0$ and any β_0 fulfilling $1 \leq \|\beta_0\|_0 \leq C'n/\log(p)$.

Some generic selection schemes

Cross-Validation

- Hold-out
- V-fold CV
- Leave-q-out

Penalized empirical lost

- Penalized log-likelihood (AIC, BIC, etc)
- Plug-in criteria (with Mallows' C_p, etc)
- Slope heuristic

Approximation versus complexity penalization

- LinSelect

LinSelect (Y. Baraud, C. G. & S. Huet)

Ingredients

- A collection $\mathcal{S}$ of linear spaces (for approximation)
- A weight function $\Delta : \mathcal{S} \to \mathbb{R}^+$ (measure of complexity)

Criterion: residuals + approximation + complexity

$$\mathrm{Crit}(\widehat{f}_\lambda) = \inf_{S \in \widehat{\mathcal{S}}} \left[\|Y - \Pi_S \widehat{f}_\lambda\|^2 + \frac{1}{2}\|\widehat{f}_\lambda - \Pi_S \widehat{f}_\lambda\|^2 + \mathrm{pen}_\Delta(S)\widehat{\sigma}_S^2 \right]$$

where

- $\widehat{\mathcal{S}} \subset \mathcal{S}$, possibly data-dependent,
- Π_S orthogonal projector onto S,
- $\mathrm{pen}_\Delta(S) \asymp \dim(S) \vee 2\Delta(S)$ when $\dim(S) \vee 2\Delta(S) \leq 2n/3$,
- $\widehat{\sigma}_S^2 = \frac{\|Y - \Pi_S Y\|_2^2}{n - \dim(S)}$.

Non-asymptotic risk bound

Assumptions

1. $1 \leq \dim(S) \vee 2\Delta(S) \leq 2n/3$ for all $S \in \mathcal{S}$,
2. $\sum_{S \in \mathcal{S}} e^{-\Delta(S)} \leq 1$.

Theorem (Y. Baraud, C.G., S. Huet)

$$\mathbb{E}\left[\|f - \widehat{f}_{\widehat{\lambda}}\|^2\right] \leq$$

$$C\ \mathbb{E}\left[\inf_{\lambda \in \Lambda}\left\{\|f - \widehat{f}_\lambda\|^2 + \inf_{S \in \widehat{\mathcal{S}}}\left\{\|\widehat{f}_\lambda - \Pi_S \widehat{f}_\lambda\|^2 + [\dim(S) \vee \Delta(S)]\sigma^2\right\}\right\}\right]$$

The bound also holds in deviation.

Sparse linear regression

Instantiation of LinSelect

Estimators

Linear regressor: $\left\{\widehat{f}_\lambda = X\widehat{\beta}_\lambda : \ \lambda \in \Lambda\right\}$.
(e.g. Lasso, Exponential-Weighting, etc)

Approximation and complexity

- $\mathcal{S} = \left\{\text{range}(\mathbf{X}_{\mathcal{J}}) : \ \mathcal{J} \subset \{1, \ldots, p\}, \ 1 \leq |\mathcal{J}| \leq n/(3\log p)\right\}$
- $\Delta(S) \ = \ \log \binom{p}{\dim(S)} + \log(\dim(S)) \ \approx \ \dim(S)\log(p)$.

Subcollection $\widehat{\mathcal{S}}$

We set $\widehat{S}_\lambda = \text{range}\left(\mathbf{X}_{\text{supp}(\hat{\beta}_\lambda)}\right)$ and define

$$\widehat{\mathcal{S}} = \left\{\widehat{S}_\lambda, \ \lambda \in \widehat{\Lambda}\right\}, \quad \text{where } \widehat{\Lambda} = \left\{\lambda \in \Lambda \ : \ \widehat{S}_\lambda \in \mathcal{S}\right\}.$$

Case of the Lasso estimators

Lasso estimators

$$\widehat{\beta}_\lambda = \operatorname*{argmin}_{\beta} \left\{ \| Y - \mathbf{X}\beta \|^2 + 2\lambda \|\beta\|_1 \right\}, \quad \lambda > 0$$

Parameter tuning: theory

For $\mathbf{X}$ with columns normalized to 1

$$\lambda \asymp \sigma \sqrt{2 \log(p)}$$

Parameter tuning: practice

- V-fold CV
- BIC criterion

Recent criterions pivotal with respect to the variance

- ℓ_1**-penalized log-likelihood.** (Stadler, Buhlmann, van de Geer)

$$\widehat{\beta}_\lambda^{LL}, \widehat{\sigma}_\lambda^{LL} := \underset{\beta \in \mathbb{R}^p, \sigma' > 0}{\operatorname{argmin}} \left[n \log(\sigma') + \frac{\|Y - \mathbf{X}\beta\|_2^2}{2\sigma'^2} + \lambda \frac{\|\beta\|_1}{\sigma'} \right].$$

- ℓ_1**-penalized Huber's loss.** (Belloni *et al.*, Antoniadis)

$$\widehat{\beta}_\lambda^{SR}, \widehat{\sigma}_\lambda^{SR} := \underset{\beta \in \mathbb{R}^p, \sigma' > 0}{\operatorname{argmin}} \left[\frac{n\sigma'}{2} + \frac{\|Y - \mathbf{X}\beta\|_2^2}{2\sigma'} + \lambda \|\beta\|_1 \right].$$

Equivalent to Square-Root Lasso (introduced before)

$$\widehat{\beta}_\lambda^{SR} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left[\sqrt{\|Y - \mathbf{X}\beta\|_2^2} + \frac{\lambda}{\sqrt{n}} \|\beta\|_1 \right].$$

Sun & Zhang : optimization with a single LARS-call

The compatibility constant

$$\kappa[\xi, T] = \min_{u \in \mathcal{C}(\xi, T)} \left\{ |T|^{1/2} \|\mathbf{X}u\|_2 / \|u_T\|_1 \right\},$$

where $\mathcal{C}(\xi, T) = \{u : \ \|u_{T^c}\|_1 < \xi \|u_T\|_1\}$.

Restricted eigenvalue

For $k^* = n/(3\log(p))$ we set $\phi_* = \sup\{\|Xu\|_2/\|u\|_2 : u \ k^*\text{-sparse}\}$

Theorem for Square-Root Lasso (Sun & Zhang)

For $\lambda = 2\sqrt{2\log(p)}$, if we assume that

- $\|\beta_0\|_0 \le C_1 \, \kappa^2[4, \mathrm{supp}(\beta_0)] \times \frac{n}{\log(p)}$,

then, with high probability,

$$\|\mathbf{X}(\widehat{\beta} - \beta_0)\|_2^2 \quad \le \quad \inf_{\beta \neq 0} \left\{ \|\mathbf{X}(\beta_0 - \beta)\|_2^2 + C_2 \frac{\|\beta\|_0 \log(p)}{\kappa^2[4, \mathrm{supp}(\beta)]} \sigma^2 \right\}.$$

The compatibility constant

$$\kappa[\xi, T] = \min_{u \in \mathcal{C}(\xi,T)} \left\{ |T|^{1/2} \|\mathbf{X}u\|_2 / \|u_T\|_1 \right\},$$

where $\mathcal{C}(\xi, T) = \{u : \ \|u_{T^c}\|_1 < \xi \|u_T\|_1\}$.

Restricted eigenvalue

For $k^* = n/(3\log(p))$ we set $\phi_* = \sup \{\|Xu\|_2/\|u\|_2 : u \ k^*\text{-sparse}\}$

Theorem for LinSelect Lasso

If we assume that

- $\|\beta_0\|_0 \leq C_1 \, \kappa^2[4, \mathrm{supp}(\beta_0)] \times \frac{n}{\phi_* \log(p)}$,

then, with high probability,

$$\|\mathbf{X}(\widehat{\beta} - \beta_0)\|_2^2 \ \leq \ C \inf_{\beta \neq 0} \left\{ \|\mathbf{X}(\beta_0 - \beta)\|_2^2 + C_2 \frac{\|\beta\|_0 \log(p)}{\phi_* \kappa^2[4, \mathrm{supp}(\beta)]} \sigma^2 \right\}.$$

Numerical experiments (1/2)

Tuning the Lasso

- 165 examples extracted from the literature
- each example e is evaluated on the basis of 400 runs

Comparison to the oracle $\widehat{\beta}_{\lambda^*}$

procedure	quantiles				
	0%	50%	75%	90%	95%
Lasso 10-fold CV	1.03	1.11	1.15	1.19	1.24
Lasso LinSelect	0.97	1.03	1.06	1.19	2.52
Square-Root Lasso	1.32	2.61	3.37	11.2	17

For each procedure ℓ, quantiles of $\mathcal{R}\left[\widehat{\beta}_{\hat{\lambda}_\ell};\beta_0\right]/\mathcal{R}\left[\widehat{\beta}_{\lambda^*};\beta_0\right]$, for $e=1,\ldots,165$.

Numerical experiments (2/2)

Computation time

n	p	10-fold CV	LinSelect	Square-Root
100	100	4 s	0.21 s	0.18 s
100	500	4.8 s	0.43 s	0.4 s
500	500	300 s	11 s	6.3 s

Packages:

- `enet` for 10-fold CV and LinSelect
- `lars` for Square-Root Lasso (procedure of Sun & Zhang)

Non-parametric regression

An important class of estimators

Linear estimators : $\widehat{f}_\lambda = A_\lambda Y$ with $A_\lambda \in \mathbb{R}^{n \times n}$

- spline smoothing or kernel ridge estimators with smoothing parameter $\lambda \in \mathbb{R}^+$
- Nadaraya estimators A_λ with smoothing parameter $\lambda \in \mathbb{R}^+$
- λ-nearest neighbors, $\lambda \in \{1, \ldots, k\}$
- L^2-basis projection (on the λ first elements)
- etc

Selection criterions (with σ^2 unknown)

- Cross-Validation schemes (including GCV)
- Mallows' C_L + plug-in / slope heuristic
- LinSelect

An important class of estimators

Linear estimators : $\widehat{f}_\lambda = A_\lambda Y$ with $A_\lambda \in \mathbb{R}^{n \times n}$

- spline smoothing or kernel ridge estimators with smoothing parameter $\lambda \in \mathbb{R}^+$
- Nadaraya estimators A_λ with smoothing parameter $\lambda \in \mathbb{R}^+$
- λ-nearest neighbors, $\lambda \in \{1, \dots, k\}$
- L^2-basis projection (on the λ first elements)
- etc

Selection criterions (with σ^2 unknown)

- Cross-Validation schemes (including GCV)
- Mallows' C_L + plug-in / slope heuristic
- LinSelect

Slope heuristic (Arlot & Bach)

Procedure for $\widehat{f}_\lambda = A_\lambda Y$

1. compute $\widehat{\lambda}_0(\sigma') = \operatorname{argmin}_\lambda \left\{ \| Y - \widehat{f}_\lambda \|^2 + \sigma' \mathrm{Tr}(2A_\lambda - A_\lambda^* A_\lambda) \right\}$
2. select $\widehat{\sigma}$ such that $\mathrm{Tr}(A_{\hat{\lambda}_0(\hat{\sigma})}) \in [n/10, n/3]$
3. select $\widehat{\lambda} = \operatorname{argmin}_\lambda \left\{ \| Y - \widehat{f}_\lambda \|^2 + 2\,\widehat{\sigma}^2\, \mathrm{Tr}(A_\lambda) \right\}$.

Main assumptions

- $A_\lambda \approx$ shrinkage or "averaging" matrix (covers all classics)
- **Bias assumption :**
  $\exists \lambda_1, \quad \mathrm{Tr}(A_{\lambda_1}) \leq \sqrt{n} \quad \text{and} \quad \|(I - A_{\lambda_1}) f\|^2 \leq \sigma^2 \sqrt{n \log(n)}$

Theorem (Arlot & Bach)

With high proba: $\|\widehat{f}_{\hat{\lambda}} - f\|^2 \leq (1+\varepsilon) \inf_\lambda \|\widehat{f}_\lambda - f\|^2 + C\, \varepsilon^{-1} \log(n) \sigma^2$

LinSelect

Approximation spaces

$\widehat{\mathcal{S}} = \bigcup_\lambda \left\{ S_\lambda^1, \ldots, S_\lambda^{n/2} \right\}$ where S_λ^k is spanned by "the k last" right-singular vectors of $A_\lambda^+ - \bar{\Pi}_\lambda$: range$(A_\lambda) \to$ range(A_λ^*), where

- A_λ^+ is the inverse of the of A_λ to range$(A_\lambda^*) \to$ range(A_λ)
- $\bar{\Pi}_\lambda$ is induced by the projection onto range(A_λ^*)

Weight

$\Delta(S) = \beta\big(1 + \dim(S)\big)$ with $\beta > 0$ such that $\sum_S e^{-\Delta(S)} \leq 1$.

Corollary

When $\sigma_{n/2}(A_\lambda^+ - \bar{\Pi}_\lambda) \geq 1/2$ for all $\lambda \in \Lambda$, we have

$$\mathbb{E}\left[\|\widehat{f} - f\|^2\right] \leq C \inf_{\lambda \in \Lambda} \mathbb{E}\left[\|\widehat{f}_\lambda - f\|^2\right]$$

LinSelect

Approximation spaces

$\widehat{\mathcal{S}} = \bigcup_\lambda \left\{S_\lambda^1, \ldots, S_\lambda^{n/2}\right\}$ where S_λ^k is spanned by "the k last" right-singular vectors of $A_\lambda^+ - \bar{\Pi}_\lambda$: range$(A_\lambda) \to$ range(A_λ^*),

Remark: when A_λ is symmetric positive definite, S_λ^k is spanned by "the k first" eigenvectors of A_λ.

Weight

$\Delta(S) = \beta\big(1 + \dim(S)\big)$ with $\beta > 0$ such that $\sum_S e^{-\Delta(S)} \leq 1$.

Corollary

When $\sigma_{n/2}(A_\lambda^+ - \bar{\Pi}_\lambda) \geq 1/2$ for all $\lambda \in \Lambda$, we have

$$\mathbb{E}\left[\|\widehat{f} - f\|^2\right] \leq C \inf_{\lambda \in \Lambda} \mathbb{E}\left[\|\widehat{f}_\lambda - f\|^2\right]$$

LinSelect

Approximation spaces

$\widehat{\mathcal{S}} = \bigcup_\lambda \left\{ S_\lambda^1, \ldots, S_\lambda^{n/2} \right\}$ where S_λ^k is spanned by "the k last" right-singular vectors of $A_\lambda^+ - \bar{\Pi}_\lambda :$ range$(A_\lambda) \to$ range(A_λ^*),

Remark: when A_λ is symmetric positive definite, S_λ^k is spanned by "the k first" eigenvectors of A_λ.

Weight

$\Delta(S) = \beta\big(1 + \dim(S)\big)$ with $\beta > 0$ such that $\sum_S e^{-\Delta(S)} \leq 1$.

Corollary

When $\sigma_{n/2}(A_\lambda^+ - \bar{\Pi}_\lambda) \geq 1/2$ for all $\lambda \in \Lambda$, we have

$$\mathbb{E}\left[\|\widehat{f} - f\|^2\right] \leq C \inf_{\lambda \in \Lambda} \mathbb{E}\left[\|\widehat{f}_\lambda - f\|^2\right]$$

LinSelect

Approximation spaces

$\widehat{\mathcal{S}} = \bigcup_\lambda \left\{S_\lambda^1, \ldots, S_\lambda^{n/2}\right\}$ where S_λ^k is spanned by "the k last" right-singular vectors of $A_\lambda^+ - \bar{\Pi}_\lambda$: range$(A_\lambda) \to$ range(A_λ^*),

Remark: when A_λ is symmetric positive definite, S_λ^k is spanned by "the k first" eigenvectors of A_λ.

Weight

$\Delta(S) = \beta\big(1 + \dim(S)\big)$ with $\beta > 0$ such that $\sum_S e^{-\Delta(S)} \leq 1$.

Corollary

When $\sigma_{n/2}(A_\lambda^+ - \bar{\Pi}_\lambda) \geq 1/2$ for all $\lambda \in \Lambda$, we have with high proba

$$\|\widehat{f} - f\|^2 \leq C \inf_{\lambda \in \Lambda} \|\widehat{f}_\lambda - f\|^2 + \log(n)\sigma^2$$

A review

High-dimensional regression with unknown variance
C.G., S. Huet & N. Verzelen
arXiv:1109.5587

(including coordinate-sparsity, group-sparsity, variation-sparsity and multivariate regression)