Optimal Aggregation of Affine Estimators

Arnak Dalalyan, École des Ponts ParisTech
Joseph Salmon, Duke University

COLT 2011

Introduction

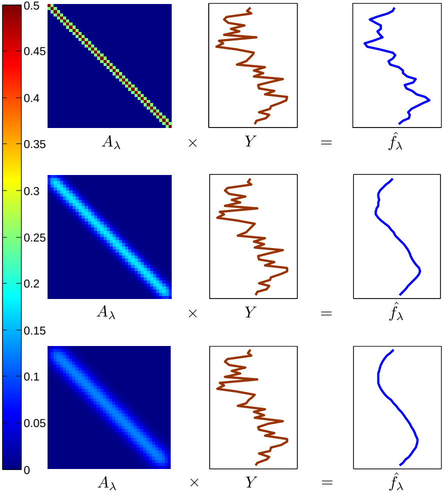
Motivations

- Theoretical : oracle inequalities (high dimension, sparsity), Adaptation in the regression model
- Applications : image processing, genetics, inverse problems (derivative estimation, deconvolution with a known kernel, tomography), etc.

Underlying Heuristic

- Aggregating/mixing estimators can be more stable than selecting only one estimator

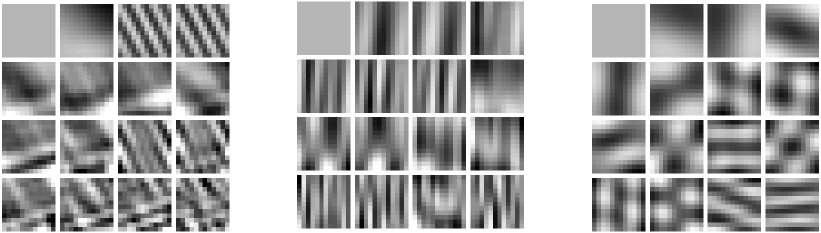
Motivations : doing as good as the best filter



$Y \in \mathbb{R}^n$: noisy signal
$\hat{f}_\lambda$: estimated signal
A_λ : convolution/filter/kernel matrix indexed by some smoothing parameter (bandwidth) λ in a family Λ
$\mathcal{F}_\Lambda$: family of estimators

Motivations : doing as good as the best dictionary approximation

Image denoising with patches



Dictionary (Dictionaries ?)

Estimate an image/patch Y by $\hat{f}_\lambda = f_\lambda = \sum_{j=1}^{M} \lambda_j \varphi_j$, for some dictionary/frame/orthonormal basis $\{\varphi_j, j = 1, \cdots, M\}$
$\mathcal{F}_\Lambda = \mathrm{Span}(\varphi_1, \cdots, \varphi_M)$ and the $\lambda = (\lambda_1, \cdots, \lambda_M)$ are the coefficients

Penalization Methods

Assume $\hat{f}_\lambda = f_\lambda = \sum_{j=1}^{M} \lambda_j \varphi_j$, for some features $\varphi_j \in \mathbb{R}^n$ and

$\hat{r}_\lambda = \| Y - \hat{f}_\lambda \|_n^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_{\lambda,i})^2$: empirical quadratic risk

Penalization Methods

$$\hat{f}^{\text{Pen}} = f_{\hat{\lambda}}, \quad \text{where} \quad \hat{\lambda} = \arg\min_{\lambda \in \Lambda} \Big(\underbrace{\hat{r}_\lambda}_{\text{data-fitting}} + \underbrace{\text{Pen}(\lambda)}_{\text{regularization}} \Big)$$

- $\text{Pen}(\lambda) = \beta \|\lambda\|_2^2$: Ridge Tikhonov [43]
- $\text{Pen}(\lambda) = \beta \|\lambda\|_0$: AIC,BIC Akaike [74], Schwarz [78]
- $\text{Pen}(\lambda) = \beta \|\lambda\|_1$: LASSO Tibshirani [96]

Rem 1 : β smoothing parameter

Rem 2 : possible blocks/mixture versions (*eg.* Elastic Net)

Rem 3 : one usually uses only one estimate in the end : $f_{\hat{\lambda}}$

Mixing classical filtering and dictionary learning

- Y : noisy vector/patch of pixels intensities, f the true one.
- Classical filtering : estimate f by AY, A convolution matrix.
  - Sharp oracle inequality for mixing estimators of the form AY with A projection matrix (Countable family) Leung and Barron [06]
- Dictionary learning : estimate f combining features b that are essentially independent of Y.
  - Sharp oracle inequality for mixing estimators built on an independent sample Dalalyan and Tsybakov [07,08]
- Goal : extending those results to aggregate estimates of the form $AY + b$ with A and b independent of Y.

NP Estimation vs. Aggregation

	Available	Non Available	Target
NP Estimation	Y	f	the best estimator
Aggregation	$Y, \mathcal{F}_\Lambda$	f	an estimator (almost) as good as the best in the family

Advantage : no need to evaluate the approximation term

Notation and model

Gaussian Heteroscedastic Model

$$Y_i = f_i + \sigma_i \varepsilon_i, \quad i = 1, \cdots, n \qquad (\star)$$

ε_i i.i.d $\mathcal{N}(0,1)$ and $\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_n^2)$ (Σ known)

- Rem 1 : $f_i = f(x_i), (x_i)_{i=1,\cdots,n}$ fixed design (cf. pixels)
- Rem 2 : $\Sigma = \sigma^2 I_n$, homoscedastic model

Goal : estimate f by $\hat{f}$, with a small (quadratic) risk

$$r = \mathbb{E}\left(\left\| f - \hat{f} \right\|_n^2 \right) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} (f_i - \hat{f}_i)^2 \right)$$

Rem : link with inverse problems wit known operator Cavalier [08]

Aggregation of Estimators and Oracle Inequalities

Family of « pre-estimators » : $\mathcal{F}_\Lambda = \{\hat{f}_\lambda \in \mathbb{R}^n, \lambda \in \Lambda\}, \Lambda \subset \mathbb{R}^M$
Goal : proving an oracle inequality for an estimator $\hat{f}_{aggr}$

Oracle Inequality Nemirovski [00]

$$\mathbb{E}\|\hat{f}_{agr} - f\|_n^2 \leq C_n \inf_{\lambda \in \Lambda} \mathbb{E}\|\hat{f}_\lambda - f\|_n^2 + R_{n,\Lambda}$$

- An **Oracle** is any $\hat{f}_{\lambda^*}$ s.t. $\lambda^* \in \arg\min_{\lambda \in \mathcal{F}_\lambda} \mathbb{E}\|\hat{f}_\lambda - f\|_n^2$
- $C_n \geq 1$. When $C_n = 1$: the inequality is said **Sharp**
- $R_{n,\Lambda} \overset{n\to\infty}{\longrightarrow} 0$: price to pay for not knowing the Oracle, depends on the complexity of Λ and on the noise intensity

Rem 1 : $\hat{f}_{agr}$ might not be in $\mathcal{F}_\Lambda$
Rem 2 : Optimality (lower bound) for some sets Λ Tsybakov [03]

EWA : classical point of view

EWA/Gibbs Measure

$$\boxed{\hat{\pi}^{\mathsf{EWA}}(d\lambda) \propto \exp(-n\hat{r}_\lambda/\beta)\pi(d\lambda)}$$

- π : prior over Λ
- $\hat{\pi}^{\mathsf{EWA}}$: posterior over Λ
- β : smoothing parameter/temperature
- $\hat{r}_\lambda$: unbiased risk estimate $\mathbb{E}(\hat{r}_\lambda) = \mathbb{E}\|\hat{f}_\lambda - f\|_n^2 = r_\lambda$

Posterior expectation : $\boxed{\hat{f}^{\mathsf{EWA}} = \int_\Lambda \hat{f}_\lambda \hat{\pi}^{\mathsf{EWA}}(d\lambda)}$

Rem 1 : -if $\beta \to 0$, $\hat{f}^{\mathsf{EWA}} \to \hat{f}_{\lambda^*}$ with $\lambda^* = \arg\min_{\lambda \in \Lambda} \hat{r}_\lambda$

-if $\beta \to \infty$, $\hat{f}^{\mathsf{EWA}} \to \int_\Lambda \hat{f}_\lambda \pi(d\lambda)$

Rem 2 : the unbiased risk estimate $\hat{r}_\lambda$ relies on Stein's Lemma Stein [81]

EWA : Penalty point of view

- Extension : enlarge the parameter space and adapt the penalty
- Parameter space : $\mathcal{P}_\Lambda = \{p : \text{probability over } \Lambda\}$
- Extended penalty : $\hat{f}^{\text{Pen}} = \int_\Lambda \hat{f}_\lambda \hat{\pi}^{\text{Pen}}(d\lambda)$ with

$$\hat{\pi}^{\text{Pen}} = \arg\min_{p \in \mathcal{P}_\Lambda} \left(\int_\Lambda \hat{r}_\lambda p(d\lambda) + \int_\Lambda \text{Pen}(\lambda) p(d\lambda) \right)$$

EWA/Kullback-Leibler penalty

$$\text{EWA} : \begin{cases} \hat{\pi}^{\text{EWA}} & = \arg\min_{p \in \mathcal{P}_\Lambda} \left(\int_\Lambda \hat{r}_\lambda p(d\lambda) + \frac{\beta}{n} \mathcal{K}(p, \pi) \right) \\ \hat{f}^{\text{EWA}} & = \int_\Lambda \hat{f}_\lambda \hat{\pi}^{\text{EWA}}(d\lambda) \end{cases}$$

- π prior over Λ ; β smoothing parameter (aka « temperature »)
- $\mathcal{K}(p, \pi)$: KL-divergence between probabilities $p, \pi \in \mathcal{P}_\Lambda$,

$$\mathcal{K}(p, \pi) = \begin{cases} \int_\Lambda \log \left(\frac{dp}{d\pi}(\lambda) \right) p(d\lambda) & \text{if } p \ll \pi, \\ +\infty & \text{otherwise.} \end{cases}$$

EWA : Penalty point of view

- Extension : enlarge the parameter space and adapt the penalty
- Parameter space : $\mathcal{P}_\Lambda = \{p : \text{probability over } \Lambda\}$
- Extended penalty : $\hat{f}^{\mathrm{Pen}} = \int_\Lambda \hat{f}_\lambda \hat{\pi}^{\mathrm{Pen}}(d\lambda)$ with

$$\hat{\pi}^{\mathrm{Pen}} = \arg\min_{p \in \mathcal{P}_\Lambda} \left(\int_\Lambda \hat{r}_\lambda p(d\lambda) + \int_\Lambda \mathrm{Pen}(\lambda) p(d\lambda) \right)$$

EWA/Kullback-Leibler penalty

$$\mathrm{EWA} : \begin{cases} \hat{\pi}^{\mathsf{EWA}} & = \arg\min_{p \in \mathcal{P}_\Lambda} \left(\int_\Lambda \hat{r}_\lambda p(d\lambda) + \frac{\beta}{n} \mathcal{K}(p, \pi) \right) \\ \hat{f}^{\mathsf{EWA}} & = \int_\Lambda \hat{f}_\lambda \hat{\pi}^{\mathsf{EWA}}(d\lambda) \end{cases}$$

- π prior over Λ ; β smoothing parameter (aka « temperature »)
- $\mathcal{K}(p, \pi)$: KL-divergence between probabilities $p, \pi \in \mathcal{P}_\Lambda$,

$$\mathcal{K}(p, \pi) = \begin{cases} \int_\Lambda \log \left(\frac{dp}{d\pi}(\lambda) \right) p(d\lambda) & \text{if } p \ll \pi, \\ +\infty & \text{otherwise.} \end{cases}$$

Affine estimators

Affine estimators

$$\boxed{\hat{f}_\lambda = A_\lambda Y + b_\lambda}$$

- A_λ : $n \times n$ matrix ; b_λ : deterministic vector in $\mathbb{R}^n$
- A_λ , b_λ : independent of Y
- Λ : possibly non-countable

Constant case : $A_\lambda = 0$, $\hat{f}_\lambda = b_\lambda$

$\{\varphi_1, \cdots, \varphi_M\}$ is a finite « dictionary »of features

- $\mathcal{F}_\Lambda = \{\varphi_1, \cdots, \varphi_M\}$ finite family
- $\mathcal{F}_\Lambda = \mathrm{conv}(\varphi_1, \cdots, \varphi_M)$ convex combinations
- $\mathcal{F}_\Lambda = \mathrm{Span}(\varphi_1, \cdots, \varphi_M)$ linear combinations
- $\mathcal{F}_\Lambda = \mathrm{Span}_S(\varphi_1, \cdots, \varphi_M)$ S-sparse combinations

Lower bounds : Tsybakov [03], Bunea et al. [07] , Lounici [07]

Linear case : $\hat{f}_\lambda = A_\lambda Y \quad (b_\lambda = 0)$

Ordinary Least Squares

$\{\mathcal{S}_\lambda : \lambda \in \Lambda\}$ family of subspaces of $\mathbb{R}^n$ A_λ : orthogonal projectors over $\mathcal{S}_\lambda$ Leung and Barron [06], Alquier and Lounici [10], Rigollet and Tsybakov [11]

Diagonal Matrices : $A_\lambda = \mathrm{diag}(a_1, \dots, a_n)$

- Ordered projections : $a_k = \mathbb{1}_{(k \le \lambda)}$ for λ integer, ie. $\Lambda = \{1, \dots, n\}$
- Pinsker's Filter : $a_k = \left(1 - \frac{k^\alpha}{w}\right)_+$, with $x_+ = \max(x, 0)$ and $w, \alpha > 0$, i.e., $\Lambda = (\mathbb{R}_+^*)^2$
- ...

Linear case : $\hat{f}_\lambda = A_\lambda Y \quad (b_\lambda = 0)$

Ordinary Least Squares

$\{\mathcal{S}_\lambda : \lambda \in \Lambda\}$ family of subspaces of $\mathbb{R}^n$ A_λ : orthogonal projectors over $\mathcal{S}_\lambda$ Leung and Barron [06], Alquier and Lounici [10], Rigollet and Tsybakov [11]

Diagonal Matrices : $A_\lambda = \mathrm{diag}(a_1, \dots, a_n)$

- Ordered projections : $a_k = \mathbb{1}_{(k \leq \lambda)}$ for λ integer, ie. $\Lambda = \{1, \dots, n\}$
- Pinsker's Filter : $a_k = \left(1 - \frac{k^\alpha}{w}\right)_+$, with $x_+ = \max(x, 0)$ and $w, \alpha > 0$, i.e., $\Lambda = (\mathbb{R}_+^*)^2$
- ...

Main theorem conditions

$$\boxed{\hat{f}_\lambda = A_\lambda Y + b_\lambda}$$

Condition C_1

- Matrices A_λ : orthogonal projections ($A_\lambda^2 = A_\lambda^\top = A_\lambda$)
- Vectors b_λ : $A_\lambda b_\lambda = 0$

Example : A_λ projectors on subspaces Leung and Barron [06]

Condition C_2

- Matrices A_λ : symmetric, positive semi-definite
- $A_\lambda A_{\lambda'} = A_{\lambda'} A_\lambda, \forall \lambda, \lambda' \in \Lambda$ and $A_\lambda \Sigma = \Sigma A_\lambda, \forall \lambda \in \Lambda$
- Vectors b_λ : $A_{\lambda'} b_\lambda = 0, \forall \lambda, \lambda' \in \Lambda$

Example : two-blocks James-Stein shrinking estimators Leung [04]

Main theorem conditions

$$\boxed{\hat{f}_\lambda = A_\lambda Y + b_\lambda}$$

Condition C_1

- Matrices A_λ : orthogonal projections ($A_\lambda^2 = A_\lambda^\top = A_\lambda$)
- Vectors b_λ : $A_\lambda b_\lambda = 0$

Example : A_λ projectors on subspaces Leung and Barron [06]

Condition C_2

- Matrices A_λ : symmetric, positive semi-definite
- $A_\lambda A_{\lambda'} = A_{\lambda'} A_\lambda, \forall \lambda, \lambda' \in \Lambda$ and $A_\lambda \Sigma = \Sigma A_\lambda, \forall \lambda \in \Lambda$
- Vectors b_λ : $A_{\lambda'} b_\lambda = 0, \forall \lambda, \lambda' \in \Lambda$

Example : two-blocks James-Stein shrinking estimators Leung [04]

Main Theorem

PAC (EAC) - Bayesian Bound

If $\mathbf{C}_1$ or $\mathbf{C}_2$ is satisfied, then for any prior π, $\hat{f}^{\text{EWA}}$ satisfies π :

$$\mathbb{E}(\|\hat{f}^{\text{EWA}} - f\|_n^2) \leq \inf_{p \in \mathcal{P}_\Lambda} \left(\int_\Lambda \mathbb{E}\|\hat{f}_\lambda - f\|_n^2 \, p(d\lambda) + \frac{\beta}{n} \, \mathcal{K}(p, \pi) \right)$$

where $\beta \geq 4 \max_{i=1,\dots,n} \sigma_i^2$ under $\mathbf{C}_1$

$\beta \geq 8 \max_{i=1,\dots,n} \sigma_i^2$ under $\mathbf{C}_2$

with $\mathcal{K}(p, \pi)$ the KL divergence between p and π

Corollary : finite case

Oracle Inequality : $\Lambda = [\![1, M]\!]$, π uniform

If $\mathbf{C}_1$ or $\mathbf{C}_2$ is satisfied, and if π is uniform on $[\![1, M]\!]$, then

$$\mathbb{E}(\|\hat{f}^{\text{EWA}} - f\|_n^2) \leq \inf_{\lambda \in [\![1,M]\!]} \left(\mathbb{E}\|\hat{f}_\lambda - f\|_n^2 \right) + \frac{\beta \log(M)}{n}$$

where $\beta \geq 4 \max_{i=1,\dots,n} \sigma_i^2$ under $\mathbf{C}_1$

$\beta \geq 8 \max_{i=1,\dots,n} \sigma_i^2$ under $\mathbf{C}_2$

- For $b_\lambda = 0$, it extends the result by Leung and Barron [06]
- For $A_\lambda = 0$ and if $\Sigma = \sigma I_n$: the inequality is optimal Tsybakov [03]

Minimax point of view ($\Sigma = \sigma^2 I_n$)

$\theta_k(f) = \langle f | \varphi_k \rangle_n$: Discrete Fourier coefficients
$\mathcal{D}f$: Discrete Fourier Transform of f

Sobolev Ellipsoid : $\mathcal{E}(\alpha, R) = \{f \in \mathbb{R}^n : \sum_{k=1}^{n} k^{2\alpha} \theta_k(f)^2 \leq R\}$

Pinsker's Theorem : linear estimates are minimax on ellipsoids

$$\inf_{\hat{f}} \sup_{f \in \mathcal{E}(\alpha,R)} \mathbb{E}(\|\hat{f} - f\|_n^2) \sim \inf_{A} \sup_{f \in \mathcal{F}(\alpha,R)} \mathbb{E}(\|AY - f\|_n^2)$$
$$\sim \inf_{w>0} \sup_{f \in \mathcal{E}(\alpha,R)} \mathbb{E}(\|A_{\alpha,w} Y - f\|_n^2)$$

the $\inf$ is taken among all the possible estimators $\hat{f}$ and
$A_{\alpha,w} = \mathcal{D}^{\top} \mathrm{diag}\left((1 - k^{\alpha}/w)_+ ; k = 1, \ldots, n\right) \mathcal{D}$: Pinsker's Filter

Rem : $\lambda = (\alpha, w)$ and $\Lambda = (\mathbb{R}_+^*)^2$

Corollary : Adaptation

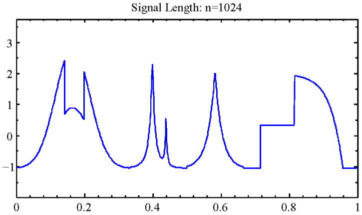
EWA on Pinsker filters : $\hat{f}_\lambda = \hat{f}_{\alpha,w} = \mathcal{D}^\top A_{\alpha,w} \mathcal{D}\, Y$ ($\mathcal{D}$: DCT),
with $A_{\alpha,w} = \mathrm{diag}((1 - \frac{k^\alpha}{w})_+, k = 1, \cdots, n)$
Choose the prior π over $\Lambda = (\mathbb{R}_+^*)^2$:

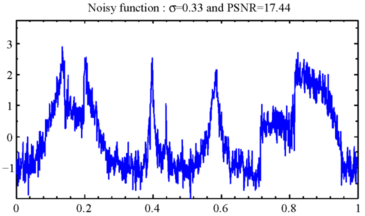
- Draw α according to an exponential distribution with parameter 1
- Knowing α, draw w according to the density $w \to \frac{2 n_\sigma^{-\alpha/(2\alpha+1)}}{\left(1+n_\sigma^{-\alpha/(2\alpha+1)} w\right)^3}$ with $n_\sigma = n/\sigma^2$

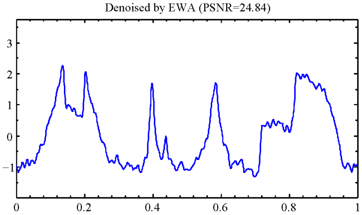
Performance

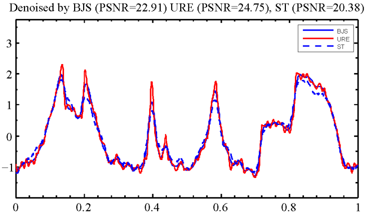
- Theoretical : adaptive in the exact minimax sense on Sobolev ellipsoids
- Practical : performance as good as other classical adaptive methods such as SURE/ Soft Thresholding Donoho and Johnstone [95] , Block James-Stein Cai [99] , empirical risk minimization Cavalier et al. [02]

1D signal experiments









Conclusion

Contributions

- Sharp oracle inequalities for affine estimators
- Adaptive results with respect to the signal smoothness
- Good experimental performance

On going work

- Weakening the assumptions for instance using a Symmetrized version of the EWA
- Extension to other type of noise

Long version of the paper and software available online :

http ://people.math.jussieu.fr/ $\sim$ salmon/

Conclusion

Contributions

- Sharp oracle inequalities for affine estimators
- Adaptive results with respect to the signal smoothness
- Good experimental performance

On going work

- Weakening the assumptions for instance using a Symmetrized version of the EWA
- Extension to other type of noise

Long version of the paper and software available online :

http ://people.math.jussieu.fr/ $\sim$ salmon/

References I

- H. Akaike.
  A new look at the statistical model identification.
  IEEE Trans. Automatic Control, AC-19 :716–723, 1974.
  System identification and time-series analysis.
- P. Alquier and K. Lounici.
  Pac-bayesian bounds for sparse regression estimation with exponential weights.
  Electron. J. Statist., 5 :127–145, 2010.
- F. Bunea, A. B. Tsybakov, and M. H. Wegkamp.
  Aggregation for Gaussian regression.
  Ann. Statist., 35(4) :1674–1697, 2007.
- T. T. Cai.
  Adaptive wavelet estimation : a block thresholding and oracle inequality approach.
  Ann. Statist., 27(3) :898–924, 1999.

References II

- L. Cavalier.
  Nonparametric statistical inverse problems.
  Inverse Problems, 24(3) :034004, 19, 2008.

- L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov.
  Oracle inequalities for inverse problems.
  Ann. Statist., 30(3) :843–874, 2002.

- D. L. Donoho and I. M. Johnstone.
  Adapting to unknown smoothness via wavelet shrinkage.
  J. Amer. Statist. Assoc., 90(432) :1200–1224, 1995.

- A. S. Dalalyan and A. B. Tsybakov.
  Aggregation by exponential weighting, sharp oracle inequalities and sparsity.
  In *COLT*, pages 97–111, 2007.

- A. S. Dalalyan and A. B. Tsybakov.
  Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity.
  Mach. Learn., 72(1-2) :39–61, 2008.

References III

- G. Leung.
  Information Theory and Mixing Least Squares Regression.
  PhD thesis, Yale University, 2004.
- K. Lounici.
  Generalized mirror averaging and D-convex aggregation.
  Math. Methods Statist., 16(3) :246–259, 2007.
- A. S. Nemirovski.
  Topics in non-parametric statistics, volume 1738 of *Lecture Notes in Math.*
  Springer, Berlin, 2000.
- Ph. Rigollet and A. B. Tsybakov.
  Exponential screening and optimal rates of sparse estimation.
  Ann. Statist., 39(2) :731–471, 2011.
- G. Schwarz.
  Estimating the dimension of a model.
  Ann. Statist., 6(2) :461–464, 1978.

References IV

- C. M. Stein.
  Estimation of the mean of a multivariate normal distribution.
  Ann. Statist., 9(6) :1135–1151, 1981.
- R. Tibshirani.
  Regression shrinkage and selection via the lasso.
  J. Roy. Statist. Soc. Ser. B, 58(1) :267–288, 1996.
- A. N. Tikhonov.
  On the stability of inverse problems.
  C. R. (Doklady) Acad. Sci. URSS (N.S.), 39 :176–179, 1943.
- A. B. Tsybakov.
  Optimal rates of aggregation.
  In *COLT*, pages 303–313, 2003.

Affine estimators and risk estimation

Stein Unbiased Risk Estimate (Gaussian Noise) Stein [81]

SURE : If $\hat{f}$ is almost everywhere differentiable in Y and $\partial_{Y_i}\hat{f}_i$ is integrable, then

$$\hat{r} = \| \boldsymbol{Y} - \hat{f}\|_n^2 + \frac{2}{n}\sum_{i=1}^{n} \partial_{Y_i}\hat{f}_i\sigma_i^2 - \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$$

is an unbaised risk estimate $\mathbb{E}(\hat{r}) = r$

SURE, Affine case : $\hat{f}_\lambda = A_\lambda Y + b_\lambda$

$$\boxed{\hat{r}_\lambda = \| \boldsymbol{Y} - \hat{f}_\lambda\|_n^2 + \frac{2}{n}\,\mathrm{Tr}(\Sigma A_\lambda) - \frac{1}{n}\,\mathrm{Tr}(\Sigma)}$$

is an unbiased risk estimate $\mathbb{E}(\|f - \hat{f}_\lambda\|_n^2) = r_\lambda$ where $\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_n^2)$