Sharp minimax and adaptive variable selection

Alexandre Tsybakov[1]

joint work with
Cristina Butucea[1,3] and Natalia Stepanova[2]

[1] ENSAE, France
[2] Carleton University, Ottawa, Canada
[3] University Paris-Est

Luminy, February 9, 2016

Overview

1. Statement of the problem
2. Non-asymptotic minimax selection bounds
3. Phase transitions
4. Adaptation to sparsity
5. Extensions and related problems

Statement of the problem

We observe

$$X_j = \theta_j + \sigma\xi_j, \quad j = 1, ..., d,$$

where $\sigma > 0$, $\xi_1, ..., \xi_d$ i.i.d. standard Gaussian r.v. and we assume that $\theta = (\theta_1, ..., \theta_d)$ belongs to

$$\begin{aligned}\Theta_d(s,a) \quad = \quad & \{\theta \in \mathbb{R}^d : \text{ there exists a set } S \subseteq \{1, \ldots, d\} \\ & \text{with } s \text{ elements , such that } |\theta_j| \geq a \text{ for all } j \in S, \\ & \text{and } \theta_j = 0 \text{ for all } j \notin S\} \,.\end{aligned}$$

Here, $a > 0$ and $s \in \{1, \ldots, d\}$ are given constants.

Variable selection problem: estimate the binary vector $\eta = (\eta_1, \ldots, \eta_d)$ where

$$\eta_j = I(\theta_j \neq 0).$$

A selector $\hat{\eta} = \hat{\eta}(X_1, \dots, X_n)$ is a binary valued estimator in $\mathbb{R}^d$:

$$\hat{\eta} = (\hat{\eta}_1, \dots, \hat{\eta}_d), \quad \hat{\eta}_j \in \{0, 1\}.$$

Hamming loss of a selector $\hat{\eta} = \hat{\eta}(X_1, \dots, X_n)$ is

$$|\hat{\eta} - \eta| := \sum_{j=1}^{d} |\hat{\eta}_j - \eta_j|.$$

Two risk measures:

- **Hamming risk**

$$E_\theta \left| \hat{\eta} - \eta \right|$$

- **Probability of wrong recovery**

$$P_\theta(S_{\hat{\eta}} \neq S(\theta))$$

where $S(\theta)$ is the support of θ (and of η), and $S_{\hat{\eta}}$ is the support of $\hat{\eta}$.

Relation between the two risk measures:

- Probability of wrong recovery is the 'Hamming risk with an indicator loss":

$$P_\theta(S_{\widehat{\eta}} \neq S(\theta)) = P_\theta(|\widehat{\eta} - \eta| \geq 1)$$

since $S_{\widehat{\eta}} = \{j : \widehat{\eta}_j = 1\}$ and $S(\theta) = \{j : \eta_j(\theta) = 1\}$.

- By Markov inequality,

$$P_\theta(S_{\widehat{\eta}} \neq S(\theta)) \leq E_\theta|\widehat{\eta} - \eta|.$$

Statistical questions:

1. **Minimax estimation w.r.t. the Hamming risk**

$$\inf_{\tilde{\eta}} \sup_{\theta\in\Theta} E_\theta \left| \tilde{\eta} - \eta \right|$$

- $\Theta = \Theta_d(s,a)$ or
- $\Theta = \Theta_d^+(s,a) = \{\theta \in \Theta_d(s,a) : \theta_j \geq 0, \forall j\}$,

$\inf_{\tilde{\eta}}$ denotes the **minimum over all selectors**.

2. **Minimax estimation w.r. to the prob. of wrong recovery**

$$\inf_{\tilde{\eta}} \sup_{\theta\in\Theta} P_\theta(S_{\tilde{\eta}} \neq S(\theta)).$$

Further statistical question:

- **Adaptive estimation w.r.t. the Hamming risk and to the Probability of wrong recovery:**

adaptation to a and s.

Further statistical question:

- **Adaptive estimation w.r.t. the Hamming risk and to the Probability of wrong recovery:**

adaptation to a and s.

- **Surprisingly:** no answers known to questions about minimax Hamming estimation, or adaptation to a and s for both risks.
- **Very rough results** available about the minimax probability of wrong recovery (in the regression/Lasso context):

If $a \geq C\sigma\sqrt{\log d}$ for $C > 0$ large enough, then there is a selector $\hat{\eta}$ such that

$$\sup_{\theta \in \Theta_d(s,a)} P_\theta(S_{\hat{\eta}} \neq S(\theta)) \to 0,$$

whereas no such selector exist if $a < c\sigma\sqrt{\log d}$, for $c > 0$ small enough (e.g., Wainwright, 2009).

Bayesian setting

- More is known about the **Bayesian setting**:
  – Genovese, Jin, Wasserman, Yao (2012) *JMLR*,
  – Ji, Jin (2012) *Ann. Statist.*

consider linear regression model with fixed and random covariates, in a Bayesian setup with $s \sim d^{1-\beta}$, for some known β in $(0, 1)$.

- There is no class $\Theta_d(s, a)$ but θ is random with independent components taking values 0 and a_d with probabilities $1 - s_d/d$ and s_d/d.
- The setting is asymptotic, $d \to \infty$.
- Hamming risk is used.
- The results are about the properties of **Exact recovery** and **Almost full recovery**:

Bayesian setting (Genovese, Jin et al., 2012)

Exact recovery (Bayesian) is possible for if there exists a selector $\hat{\eta}$ such that

$$\lim_{d\to\infty} \int E_\theta|\hat{\eta}-\eta|\mathbf{dP}_\theta = 0,$$

respectively it is impossible when

$$\liminf_{d\to\infty} \inf_{\tilde{\eta}} \int E_\theta|\tilde{\eta}-\eta|\mathbf{dP}_\theta > 0.$$

Almost full recovery (Bayesian) is possible if there exists a selector $\hat{\eta}$ such that

$$\lim_{d\to\infty} \frac{1}{s_d}\int E_\theta|\hat{\eta}-\eta|\mathbf{dP}_\theta = 0.$$

respectively, almost full recovery is impossible if

$$\liminf_{d\to\infty} \inf_{\tilde{\eta}} \frac{1}{s_d}\int E_\theta|\tilde{\eta}-\eta|\mathbf{dP}_\theta > 0.$$

Minimax setting

Exact recovery (minimax) is possible for $(\Theta_d(s_d, a_d))_{d\geq 1}$ if there exists a selector $\hat{\eta}$ such that

$$\lim_{d\to\infty} \sup_{\theta\in\Theta_d(s_d,a_d)} E_\theta|\hat{\eta}-\eta| = 0,$$

respectively it is impossible when

$$\liminf_{d\to\infty} \inf_{\tilde{\eta}} \sup_{\theta\in\Theta_d(s_d,a_d)} E_\theta|\tilde{\eta}-\eta| > 0.$$

Almost full recovery (minimax) is possible for $(\Theta_d(s_d, a_d))_{d\geq 1}$ if there exists a selector $\hat{\eta}$ such that

$$\lim_{d\to\infty} \sup_{\theta\in\Theta_d(s_d,a_d)} \frac{1}{s_d} E_\theta|\hat{\eta}-\eta| = 0.$$

respectively, almost full recovery is impossible if

$$\liminf_{d\to\infty} \inf_{\tilde{\eta}} \sup_{\theta\in\Theta_d(s_d,a_d)} \frac{1}{s_d} E_\theta|\tilde{\eta}-\eta| > 0.$$

- Ingster, Stepanova (2014) *J. Mathem Sciences*, B. and Stepanova (2015)

Gaussian white noise model, smoothness classes of θ, adaptive exact and almost full recovery.

Hamming loss was also considered e.g. in

- Neuvial, Roquain (2012) *Ann. Statist.*: oracle inequalities for multiple classification under sparsity;
- Zhang, Zhou (2015): community detection in stochastic block models. Exact recovery for minimax Hamming risk.

Non-asymptotic minimax selection bounds

Define the selector

$$\hat{\eta}_j = I(|X_j| \geq t), \quad j = 1, \dots, d, \tag{1}$$

where the threshold is defined by

$$t = \frac{a}{2} + \frac{\sigma^2}{a} \log\left(\frac{d}{s} - 1\right). \tag{2}$$

The selector is not of the form $I(|X_j| \geq C\sigma\sqrt{\log d})$!

On the positive valued set $\Theta_d^+(s, a)$, we define the selector

$$\hat{\eta}_j^+ = I(X_j \geq t), \quad j = 1, \dots, d, \tag{3}$$

with t as in (2).

Theorem 1 - Non-asymptotic minimax Hamming risk

(i) For any $a > 0$ and $s < d$ we have:

$$\sup_{\theta\in\Theta_d(s,a)} \frac{1}{s} E_\theta |\hat{\eta} - \eta| \leq 2\Psi(d, s, a),$$

$$\sup_{\theta\in\Theta_d^+(s,a)} \frac{1}{s} E_\theta |\hat{\eta}^+ - \eta| \leq \Psi_+(d, s, a).$$

(ii) Moreover,

$$\inf_{\widetilde{\eta}} \sup_{\theta\in\Theta_d^+(s,a)} \frac{1}{s} E_\theta |\widetilde{\eta} - \eta| \geq \Psi_+(d, s, a),$$

where $\inf_{\widetilde{\eta}}$ denotes the infimum over all selectors $\widetilde{\eta}$ (not necessarily separable).

The minimax constants are:

$$\Psi(d,s,a) = \left(\frac{d}{s}-1\right)\Phi\left(-\frac{a}{2\sigma}-\frac{\sigma}{a}\log\left(\frac{d}{s}-1\right)\right) + \Phi\left(-\left(\frac{a}{2\sigma}-\frac{\sigma}{a}\log\left(\frac{d}{s}-1\right)\right)_+\right),$$

$\Phi(\cdot)$ denotes the standard Gaussian cumulative distribution function, and $x_+ = \max(x, 0)$,

$$\Psi_+(d,s,a) = \left(\frac{d}{s}-1\right)\Phi\left(-\frac{a}{2\sigma}-\frac{\sigma}{a}\log\left(\frac{d}{s}-1\right)\right) + \Phi\left(-\left(\frac{a}{2\sigma}-\frac{\sigma}{a}\log\left(\frac{d}{s}-1\right)\right)\right).$$

Note that $\Psi(d,s,a) \le \Psi_+(d,s,a)$.

Proof

- **Upper bound (case of $\Theta_d^+(s,a)$):**

$$|\hat\eta^+ - \eta| = \sum_{j:\eta_j=0} I(\xi_j \geq t) + \sum_{j:\eta_j=1} I(\sigma\xi_j + \theta_j < t),$$

and

$$E\left(I\left(\sigma\xi_j + \theta_j < t\right)\right) \leq P(\xi < (t-a)/\sigma).$$

Thus, for any $\theta \in \Theta_d^+(s,a)$,

$$\frac{1}{s} E_\theta |\hat\eta^+ - \eta| \leq \left(\frac{d}{s} - 1\right) P(\xi \geq t/\sigma) + P(\xi < (t-a)/\sigma)$$

$$= \Psi_+(d,s,a).$$

Proof: Lower bound

- Reduction to separable estimators $\bar{\eta}_j$ with values in [0,1].
- Decomposition in element-wise testing problems:

$$\sup_{\theta\in\Theta_d^+(s,a)} \frac{1}{s}\sum_{j=1}^{d} E_{j,\theta_j}|\bar{\eta}_j - \eta_j| \geq$$

$$\geq \frac{d}{s}\inf_{T\in[0,1]}\left(\left(1-\frac{s}{d}\right)\mathbb{E}_0(T) + \frac{s}{d}\mathbb{E}_a(1-T)\right)$$

where $\mathbb{E}_u$ is the expectation with respect to the distribution of $X = u + \sigma\xi$ with $\xi \sim \mathcal{N}(0,1)$.

- Bayes solution is the test (φ = density of $\mathcal{N}(0,1)$)

$$T^*(X) = I\left(\frac{(s/d)\varphi_\sigma(X-a)}{(1-s/d)\varphi_\sigma(X)} > 1\right)$$

which results in the risk $\Psi_+(d,s,a)$.

Theorem 2 - Probability of wrong recovery

For any $a > 0$ and $s < d$ the selectors $\hat{\eta}$ and $\hat{\eta}^+$ with the threshold t defined in (2) satisfy

$$\sup_{\theta \in \Theta_d^+(s,a)} P_\theta(S_{\hat{\eta}^+} \neq S(\theta)) \leq s\Psi_+(d,s,a),$$

and

$$\sup_{\theta \in \Theta_d(s,a)} P_\theta(S_{\hat{\eta}} \neq S(\theta)) \leq 2s\Psi(d,s,a).$$

Furthermore,

$$\inf_{\tilde{\eta} \in \mathcal{T}} \sup_{\theta \in \Theta_d^+(s,a)} P_\theta(S_{\tilde{\eta}} \neq S(\theta)) \geq \frac{s\Psi_+(d,s,a)}{1 + s\Psi_+(d,s,a)}$$

where $\inf_{\tilde{\eta} \in \mathcal{T}}$ denotes the infimum over all **separable** selectors $\tilde{\eta}$.

Phase transitions

Theorem 3 - Necessary and sufficient conditions of almost full/exact recovery

(i) Almost full recovery is possible for $(\Theta_d(s_d, a_d))_{d\geq 1}$ if and only if

$$\Psi_+(d, s_d, a_d) \to 0 \quad \text{as } d \to \infty.$$

In this case, the selector $\hat{\eta}$ defined in (1) with threshold (2) achieves almost full recovery.

(ii) Exact recovery is possible for $(\Theta_d(s_d, a_d))_{d\geq 1}$ if and only if

$$s_d \Psi_+(d, s_d, a_d) \to 0 \quad \text{as } d \to \infty.$$

In this case, the selector $\hat{\eta}$ defined in (1) with threshold (2) achieves exact recovery.

Phase transitions

Theorem - Phase transitions Assume that $s < d/2$.

(i) If $a^2 \geq \sigma^2\Big(2\log(d/s-1)+W\Big)$ for some $W>0$, then the selector $\hat{\eta}$ defined in (1) with threshold (2) satisfies

$$\sup_{\theta\in\Theta_d(s,a)} E_\theta|\hat{\eta}-\eta| \leq (2+\sqrt{2\pi})s\,\Phi(-\Delta),$$

$$\text{where } \Delta = \frac{W}{2\sqrt{2\log(d/s-1)+W}}.$$

(ii) If $a^2 \leq \sigma^2\Big(2\log(d/s-1)+W\Big)$ for some $W>0$, then

$$\inf_{\tilde{\eta}} \sup_{\theta\in\Theta_d(s,a)} E_\theta|\tilde{\eta}-\eta| \geq s\,\Phi(-\Delta),$$

where the infimum is taken over all selectors $\tilde{\eta}$.

Assume that exact recovery is possible for the classes $(\Theta_d(s_d, a_d))_{d\geq 1}$ and $(\Theta_d^+(s_d, a_d))_{d\geq 1}$. Then, for the selectors $\hat{\eta}$ and $\hat{\eta}^+$ we have

$$
\begin{aligned}
&\lim_{d\to\infty} \sup_{\theta\in\Theta_d^+(s_d,a_d)} \frac{P_\theta(S_{\hat{\eta}^+} \neq S(\theta))}{s_d \Psi_+(d, s_d, a_d)} \\
&= \lim_{d\to\infty} \inf_{\tilde{\eta}\in\mathcal{T}} \sup_{\theta\in\Theta_d^+(s_d,a_d)} \frac{P_\theta(S_{\tilde{\eta}} \neq S(\theta))}{s_d \Psi_+(d, s_d, a_d)} = 1,
\end{aligned}
$$

and

$$
\limsup_{d\to\infty} \sup_{\theta\in\Theta_d(s_d,a_d)} \frac{P_\theta(S_{\hat{\eta}} \neq S(\theta))}{s_d \Psi_+(d, s_d, a_d)} \leq 2,
$$

$$
\liminf_{d\to\infty} \inf_{\tilde{\eta}\in\mathcal{T}} \sup_{\theta\in\Theta_d(s_d,a_d)} \frac{P_\theta(S_{\tilde{\eta}} \neq S(\theta))}{s_d \Psi_+(d, s_d, a_d)} \geq 1.
$$

Almost full recovery

Assume that $\limsup_{d\to\infty} s_d/d < 1/2$.

(i) If, for all d large enough,

$$a_d^2 \geq \sigma^2\Big(2\log((d-s_d)/s_d) + A_d\sqrt{2\log((d-s_d)/s_d)}\Big)$$

for an arbitrary sequence $A_d \to \infty$, as $d \to \infty$, then almost full recovery is possible.

(ii) Moreover, if there exists $A > 0$ such that for all d large enough the reverse inequality holds:

$$a_d^2 \leq \sigma^2\Big(2\log((d-s_d)/s_d) + A\sqrt{2\log((d-s_d)/s_d)}\Big)$$

then almost full recovery is impossible.

Exact recovery

Assume that $s_d \to \infty$ as $d \to \infty$, and $\limsup_{d\to\infty} s_d/d < 1/2$.

(i) If $a_d^2 \geq \sigma^2\Big(2\log((d-s_d)/s_d) + W_d\Big)$ for all d large enough, where the sequence W_d is such that

$$\liminf_{d\to\infty} \frac{W_d}{4\Big(\log(s_d) + \sqrt{\log(s_d)\log(d-s_d)}\Big)} \geq 1,$$

then exact recovery is possible;

(ii) If $a_d^2 \leq \sigma^2\Big(2\log((d-s_d)/s_d) + W_d\Big)$ for all d large enough, where the sequence W_d is such that

$$\limsup_{d\to\infty} \frac{W_d}{4\Big(\log(s_d) + \sqrt{\log(s_d)\log(d-s_d)}\Big)} < 1,$$

then exact recovery is impossible.

Adaptive exact recovery

Assume that $s_d \to \infty$ and that $\limsup_{d\to\infty} s_d/d < 1/2$.
The phase transition level for exact recovery is,

$$a_d^E = \sigma\left(\sqrt{2\log(d-s)} + \sqrt{2\log(s)}\right).$$

In particular, if $s \sim d^{1-\beta}$, then $a_d^E \sim (1+\sqrt{1-\beta})\sqrt{2\sigma^2\log d}$.

Then the optimal selector $\hat{\eta}$ has threshold

$$t = \sigma\sqrt{2\log(d-s)} \sim \sigma\sqrt{2\log d}$$

achieves exact recovery, adaptively to s.

Adaptive almost full recovery

Transition level is $a_d^{AF} \sim \sigma\sqrt{2\log(d/s-1)}$.

Consider a grid of points $\{g_1, \ldots, g_M\}$ on $\mathcal{S}_d$ where $g_j = 2^{j-1}$ and M is the maximal integer such that $g_M \le s_d^*$. For each g_m, $m = 1, ..., M$, we define a selector

$$\hat{\eta}(g_m) = (\hat{\eta}_j(g_m))_{j=1,\ldots,d} \triangleq (I(|X_j| \ge w(g_m)))_{j=1,\ldots,d}\,,$$

where

$$w(s) = \sigma\sqrt{2\log\left(\frac{d}{s}-1\right)}.$$

Note that $w(s)$ is monotonically decreasing.

Lepski-type data-driven procedure:

$$\begin{aligned}\widehat{m} \quad = \quad & \min\{m \in \{2, \dots, M\} : \\ & \sum_{j=1}^{d} I\big(w(g_k) \le |X_j| < w(g_{k-1})\big) \le \tau g_k, \\ & \text{for all } k \ge m\},\end{aligned}$$

$$\text{where } \tau = \big(\log(d/s_d^* - 1)\big)^{-\frac{1}{7}}.$$

Finally, the adaptive selector is

$$\hat{\eta}^{ad} = \hat{\eta}(g_{\widehat{m}}).$$

We assume that

$s_d \in \mathcal{S}_d \triangleq \{1, 2, \ldots, s_d^*\}$ where s_d^* is an integer such that $\dfrac{d}{s_d^*} \to \infty$

and that $s_d < d/4$ together with

$$A_d \geq 4 \left(\log \log \left(\frac{d}{s_d^*} - 1 \right) \right)^{1/2},$$

Then,

$$\lim_{d \to \infty} \sup_{\theta \in \Theta_d(s_d, a_d)} \frac{1}{s_d} E_\theta |\hat{\eta}^{ad} - \eta| = 0$$

for all sequences $(s_d, a_d)_{d \geq 1}$ such that $a_d \geq a_d^{AF}$.

Extensions and related problems

- Exact minimax results for other distributions than the Gaussian. Exponential families with monotone likelihood ratio. Caveat: no meaningful solution for the Bernoulli case.
- Two-dimensional problem: other conditions of almost full and exact recovery (Hajek, Wu and Xu, 2015). More structured subsets. Butucea and Ingster (2013) - exact recovery/ prob. of wrong recovery. Connection to detection boundary of Butucea, Ingster and Suslina (2015). Hajek, Wu and Xu (2015): Meaningful solution for the Bernoulli case.
- Exact minimax results for selection from more structured subsets?
- Sharp adaptation for the minimax Hamming risk?