Minimax risk and high-dimensional regression.

Nicolas Verzelen

INRA Montpellier SupAgro

Statlearn
March 18th 2011

Linear regression

$$\mathbf{Y} = \mathbf{X}\theta + \sigma\boldsymbol{\epsilon} \ ,$$

with

- $\mathbf{Y}$ response vector of size n.
- $\theta \in \mathbb{R}^{\boldsymbol{p}}$ is unknown.
- $\boldsymbol{\epsilon} \sim \mathcal{N}(0_{\boldsymbol{n}}, I_{\boldsymbol{n}})$. σ sometimes known.
- Design $\mathbf{X}$ of size $n \times p$

Design $\mathbf{X}$ of size $n \times p$:

1. Considered as **fixed**
2. Considered as **random** : $((\mathbf{Y}_1, \mathbf{X}_1), \ldots, (\mathbf{Y}_{\boldsymbol{n}}, \mathbf{X}_{\boldsymbol{n}}))$ are n iid. observations of the model :

$$Y = X\theta + \sigma\epsilon \text{ with } X^* \sim \mathcal{N}(0_{\boldsymbol{p}}, \Sigma) \text{ and } \epsilon \sim \mathcal{N}(0, 1).$$

Classical statistical challenges

$$\mathbf{Y} = \mathbf{X}\theta + \sigma\boldsymbol{\epsilon}$$

($\mathbf{P_1}$) : **Prediction**. Estimating a signal $\mathbb{E}[\mathbf{Y}] = \mathbf{X}\theta$.
$\neq$ prediction in a random design setting : Estimate $\mathbb{E}[Y_{\text{new}}|X_{\text{new}}]$.

($\mathbf{P_2}$) : **Linear hypothesis testing**. Testing the null hypothesis $\mathbf{H_0}$: "$\theta = 0$" against $\mathbf{H_1}$: "$\theta \neq 0$ but θ is sparse in some sense".

($\mathbf{P_3}$) : **Inverse Problem**. Estimating θ.

($\mathbf{P_4}$) : **Support estimation**. Recovering the support of θ. $\{i,\ \theta_i \neq 0\}$.
($\mathbf{P_4'}$) : **dimension reduction** . Estimating a set of covariates $\widehat{M} \subset \{1, \dots p\}$ of reasonable size which contains the support of θ with large probability. .

High dimension and sparsity

In many applications (e.g., postgenomics, fMRI), the number p of covariates is much larger than n.

Sparsity : most of the components of θ are zero.
Notation : $\Theta[k, p]$ set of k-sparse vectors.

High dimensional statistics : $k \leq n \leq p$.

- Theoretical challenges (non asymptotic analysis of procedures)
- Computational challenges : e.g. Lasso, Dantzig selector, ...

$$\widehat{\theta} := \arg\inf_{\theta'} \|\mathbf{Y} - \mathbf{X}\theta'\|_n^2 + \lambda\|\theta'\|_1$$

High dimension and sparsity

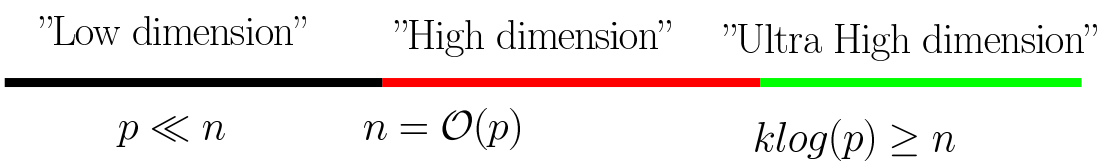
In many applications (e.g., postgenomics, fMRI), the number p of covariates is much larger than n.

Sparsity : most of the components of θ are zero.
Notation : $\Theta[k, p]$ set of k-sparse vectors.

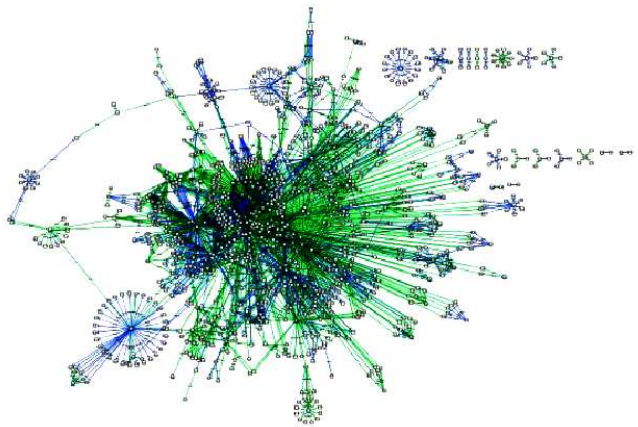
High dimensional statistics : $k \leq n \leq p$.

- Theoretical challenges (non asymptotic analysis of procedures)
- Computational challenges : e.g. Lasso, Dantzig selector, ...

$$\widehat{\theta} := \arg\inf_{\theta'} \|\mathbf{Y} - \mathbf{X}\theta'\|_n^2 + \lambda\|\theta'\|_1$$



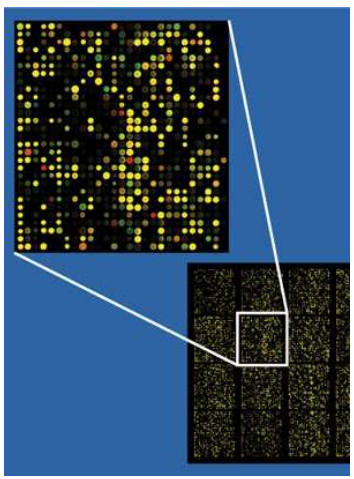
Genes of E. Coli



Complex interactions between the genes and their products that regulate the expression of the genes.

Goal : Inferring a part of the gene network using transcriptomic data

Transcriptomic = measure the gene expression levels



Analyzing dependencies patterns in the data set.

$$X_a = \sum_{b \neq a} \theta_b^a X_b + \epsilon_a \ ,$$

with $\theta_b^a \neq 0$ if and only if $a \sim b$.

Minimax properties and adaptation

Understand the structural limitations of these problems :

1. For a given problem, what is the smallest risk that one can achieve ?
2. Is it possible to get a reasonable risk bound for arbitrarily large p ?

   $\rightsquigarrow$ What can we do with $p = 5000$ genes and $n = 50$ microarray experiments ?

   $\rightsquigarrow$ $p = 200$ genes and $n = 50$ microarray experiments ?

Minimax properties and adaptation

Understand the structural limitations of these problems :

1. For a given problem, what is the smallest risk that one can achieve ?
2. Is it possible to get a reasonable risk bound for arbitrarily large p ?
   $\rightsquigarrow$ What can we do with $p = 5000$ genes and $n = 50$ microarray experiments ?
   difficult if $k \geq 4$
   $\rightsquigarrow$ $p = 200$ genes and $n = 50$ microarray experiments ? difficult if $k \geq 8$

Minimax properties and adaptation

Understand the structural limitations of these problems :

1. For a given problem, what is the smallest risk that one can achieve ?
2. Is it possible to get a reasonable risk bound for arbitrarily large p ?
   $\rightsquigarrow$ What can we do with $p = 5000$ genes and $n = 50$ microarray experiments ? difficult if $k \geq 4$
   $\rightsquigarrow$ $p = 200$ genes and $n = 50$ microarray experiments ? difficult if $k \geq 8$

Given a loss function $l(.,.)$ and an estimator $\widehat{\theta}$, the maximal risk of $\widehat{\theta}$ over $\Theta[k, p]$ is defined by

$$\sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma}[l(\widehat{\theta}, \theta)] \ .$$

The minimax risk over $\Theta[k, p]$ is

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma}[l(\widehat{\theta}, \theta)] \ .$$

Main goal : Computing the minimax risk $\Theta[k, p]$ for different loss functions associated to problems $(\mathbf{P_1} - \mathbf{P_4})$.

In practice the sparsity k is unknown and the variance σ^2 is often unknown.
Can we adapt to k ? Can we adapt to σ^2 ?

Outline

Gaussian sequence model

Particular case : $p = n$ and $\mathbf{X} = I$:

$$\mathbf{Y}_i = \theta_i + \sigma\epsilon_i, \ i = 1, \dots n$$

.

$\rightsquigarrow$ *Donoho and Johnstone (94,95), Ingster (02), Baraud (02), Donoho and Jin (04)*

Minimax risk of estimation

Proposition

$$\square_1 k \log(ep/k) \leq \inf_{\theta \in \Theta[k,p]} \|\widehat{\theta} - \theta\|_n^2/(n\sigma^2) \leq \square_2 k \log(ep/k)$$

- If the support of θ is known, the risk is k/n.
  $\rightsquigarrow$ logarithmic price because the support is unknown.
- This risk is achieved by thresholding methods (soft and hard).
- Adaptation to the sparsity is possible with thresholding techniques.

Minimax risk over worst-case designs

Goal : Estimating $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\theta$.

Loss function : $\|\mathbf{X}(\widehat{\theta} - \theta)\|_n^2/(n\sigma^2)$

Minimax risk over worst-case designs

Goal : Estimating $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\theta$.

Loss function : $\|\mathbf{X}(\widehat{\theta} - \theta)\|_n^2/(n\sigma^2)$

If the support of θ is known $\rightsquigarrow$ parametric risk k/n.

Minimax risk over worst-case designs

Goal : Estimating $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\theta$.

Loss function : $\|\mathbf{X}(\widehat{\theta} - \theta)\|_n^2/(n\sigma^2)$

If the support of θ is known $\rightsquigarrow$ parametric risk k/n.

complex dependency of the minimax risk $\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma}[\|\mathbf{X}(\widehat{\theta} - \theta)\|_n^2/(n\sigma^2)]$ on the design $\mathbf{X}$.

Goal : emphasizing the respective roles of (k, n, p).
$\rightsquigarrow$ Minimax risk uniformly over all designs $\mathbf{X}$ of size $n \times p$.

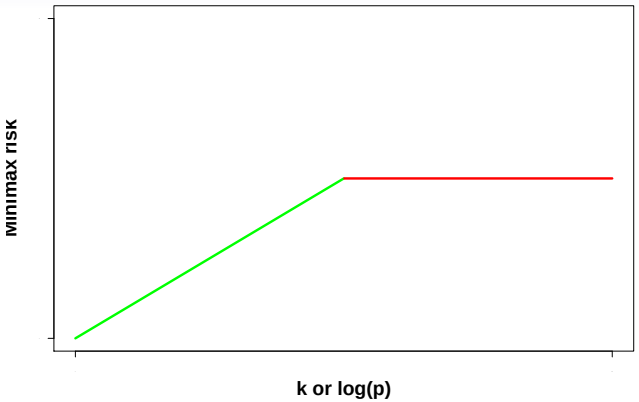
$$\mathcal{R}[k] \quad := \quad \sup_{\mathbf{X}} \inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma}[\|\mathbf{X}(\widehat{\theta} - \theta)\|_n^2/(n\sigma^2)]$$

Proposition

For all $k \leq n \wedge p$, we have

$$\square \frac{k}{n} \log(ep/k) \wedge 1 \leq \mathcal{R}[k] \leq \square' \frac{k}{n} \log(ep/k) \wedge 1$$

$$\mathcal{R}[k] \simeq \square \frac{k}{n} \log(ep/k) \wedge 1.$$



Comments :

- In reasonable dimension, "logarithmic price" if we do not known the support. $\rightsquigarrow$ analogous to the Gaussian sequence model (*Johnstone (94)*).
- In ultra-high dimension, the problem is as complex as estimating a vector in $\mathbb{R}^n$ (without any assumption).

$$\widehat{\theta}_k := \arg \inf_{\theta \in \Theta[k,p]} \|\mathbf{Y} - \mathbf{X}\theta\|_n^2 \qquad \text{si } k \leq k^*$$

$$\widehat{\theta}_n := \arg \inf_{\theta \in \mathbb{R}^n} \|\mathbf{Y} - \mathbf{X}\theta\|_n^2 \qquad \text{if } k \leq k^*$$

- Convex criteria (LASSO, Dantzig Selector) achieve these risk bounds only under restrictive hypotheses on $\mathbf{X}$.

Geometrical Interpretation

Let X be as size $n \times p$ design. Sparse eigenvalues

$$\Phi_{k,+}(\mathbf{X}) = \sup_{\theta, \|\theta\|_0 \leq k} \frac{\|\mathbf{X}\theta\|_n^2}{\|\theta\|_p^2} \qquad \Phi_{k,-}(\mathbf{X}) = \inf_{\theta, \|\theta\|_0 \leq k} \frac{\|\mathbf{X}\theta\|_n^2}{\|\theta\|_p^2}$$

Proposition

For all designs $\mathbf{X}$*, we have*

$$\mathcal{R}[k, \mathbf{X}] \quad := \quad \inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma} \left[\frac{\|\mathbf{X}(\widehat{\theta} - \theta)\|_n^2}{n\sigma^2} \right] \geq C \frac{\Phi_{2k,-}(\mathbf{X})}{\Phi_{2k,+}(\mathbf{X})} \frac{k}{n} \log\left(\frac{p}{k}\right)$$

Geometrical Interpretation

Let X be as size $n \times p$ design. Sparse eigenvalues

$$\Phi_{k,+}(\mathbf{X}) = \sup_{\theta, \|\theta\|_0 \leq k} \frac{\|\mathbf{X}\theta\|_n^2}{\|\theta\|_p^2} \qquad \Phi_{k,-}(\mathbf{X}) = \inf_{\theta, \|\theta\|_0 \leq k} \frac{\|\mathbf{X}\theta\|_n^2}{\|\theta\|_p^2}$$

Proposition

For all designs $\mathbf{X}$*, we have*

$$\mathcal{R}[k, \mathbf{X}] \quad := \quad \inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma} \left[\frac{\|\mathbf{X}(\widehat{\theta} - \theta)\|_n^2}{n\sigma^2} \right] \geq C \frac{\Phi_{2k,-}(\mathbf{X})}{\Phi_{2k,+}(\mathbf{X})} \frac{k}{n} \log\left(\frac{p}{k}\right)$$

Comments :

- If $\Phi_{2k,-}(\mathbf{X})/\Phi_{2k,+}(\mathbf{X})$ is close to 1 (condition de Riesz), then the minimax risk is of the order $k/n \log(ep/k)$ (Raskutti et al. (10)).
  Under such conditions, LASSO, Dantzig Selector, ... work well
- Lower bound based on Fano's lemma : Evaluate the "size" of the neighborhood of 0_n in $\{\mathbf{X}\theta, \theta \in \Theta[k, p]\}$.
- Corollary : Impossible to build a matrix of size $n \times p$ such that $\Phi_{2k,-}(\mathbf{X})/\Phi_{2k,+}(\mathbf{X})$ is close to 1 if $k \log(p/k)$ is large compared to n. (Baraniuk et al. 2008).

Adaptation to sparsity

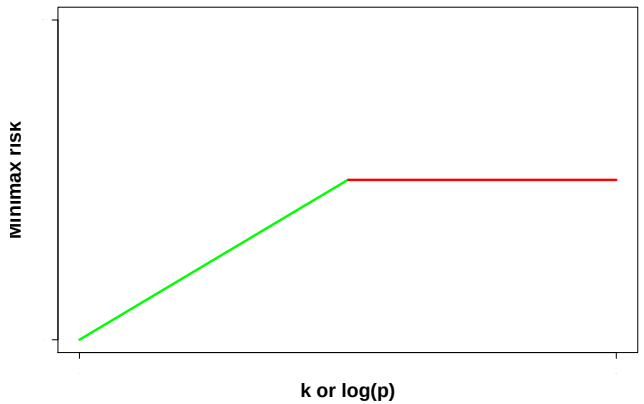
Adaptation to sparsity (penalized least-square estimators, *e.g.*, *Birgé/Massart* (01)).

$$\widehat{k} := \arg \inf_{\{k \leq k^*\} \cup \{n\}} \|\mathbf{Y} - \mathbf{X}\widehat{\theta}_k\|_n^2 + \sigma^2 \mathrm{pen}(k) \ .$$

We can take $\mathrm{pen}(k) = 3k \log(ep/k)$ for $k \leq k^*$ and $\mathrm{pen}(n) = 2n$.

- BIC and AIC underpenalize in theory and practice.
- Minimizing this criterion has a non polynomial computation burden.

Adaptation to the variance

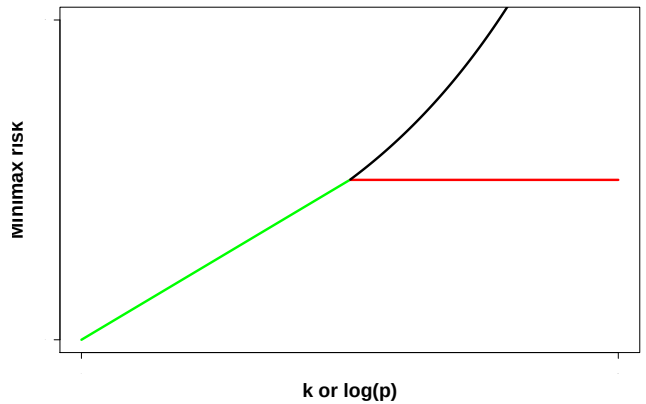


- Adaptation to the variance is possible (least-squares estimators).

$$\widehat{\theta}_k := \arg \inf_{\theta \in \Theta[k,p]} \|\mathbf{Y} - \mathbf{X}\theta\|_n^2$$

$$\widehat{\theta}_n := \arg \inf_{\theta \in \mathbb{R}^n} \|\mathbf{Y} - \mathbf{X}\theta\|_n^2$$

Adaptation to the variance



- Adaptation to the variance is possible (least-squares estimators).

$$\widehat{\theta}_k := \arg \inf_{\theta \in \Theta[k,p]} \|\mathbf{Y} - \mathbf{X}\theta\|_n^2$$

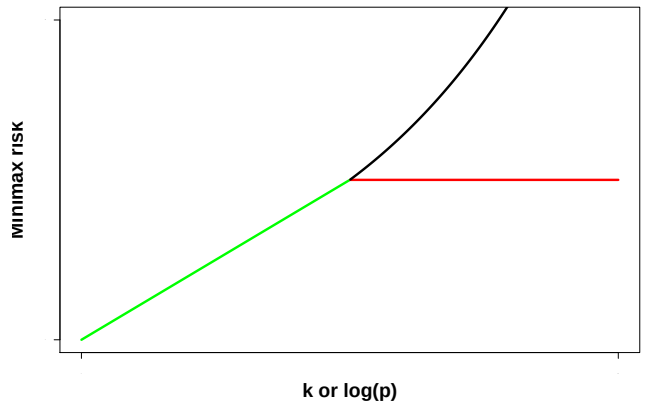
$$\widehat{\theta}_n := \arg \inf_{\theta \in \mathbb{R}^n} \|\mathbf{Y} - \mathbf{X}\theta\|_n^2$$

- Is it possible to be simultaneously adaptive to the sparsity and the variance ? *Baraud/Giraud/Huet (09)*

$$\tilde{k}_{BGH} := \arg \inf_{k \leq n/2} \|\mathbf{Y} - \mathbf{X}\widehat{\theta}_k\|_n^2 \left[1 + \psi(k)\right] \ ,$$

ψ plays the role of a penalty.

Adaptation to the variance



- Adaptation to the variance is possible (least-squares estimators).

$$\widehat{\theta}_k := \arg \inf_{\theta \in \Theta[k,p]} \|\mathbf{Y} - \mathbf{X}\theta\|_n^2$$

$$\widehat{\theta}_n := \arg \inf_{\theta \in \mathbb{R}^n} \|\mathbf{Y} - \mathbf{X}\theta\|_n^2$$

- Is it possible to be simultaneously adaptive to the sparsity and the variance ? *Baraud/Giraud/Huet (09)*

$$\tilde{k}_{BGH} := \arg \inf_{k \leq n/2} \|\mathbf{Y} - \mathbf{X}\widehat{\theta}_k\|_n^2 \left[1 + \psi(k)\right] \ ,$$

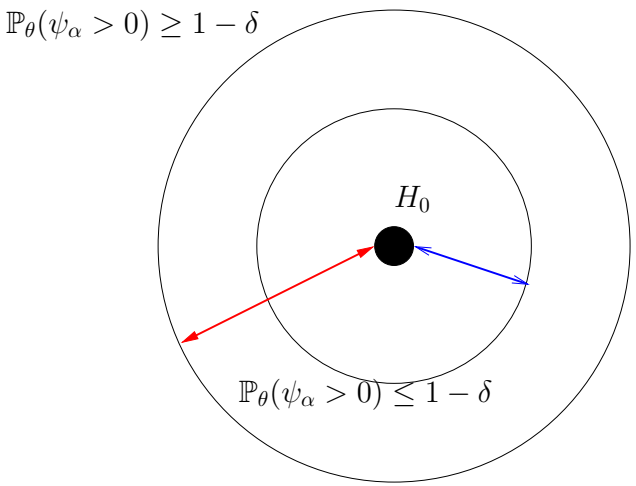
ψ plays the role of a penalty.
No, it is impossible, BGH is optimal.

Minimax separation distance

$\mathbf{H_0}$: $\theta = 0$ against $\mathbf{H_1}$: $\theta \in \Theta[k, p] \setminus \{0\}$.
Fix $\delta > 0$. ψ_α test of Level α.

Separation distance of ψ_α :

$$\rho[\psi_\alpha, k, \mathbf{X}] := \inf \left\{ \rho > 0, \inf_{\theta \in \Theta[k,p],\ \|\mathbf{X}\theta\|_n \geq \sqrt{n}\rho\sigma} \mathbb{P}_{\theta,\sigma}[\psi_\alpha = 1] \geq 1 - \delta \right\} .$$

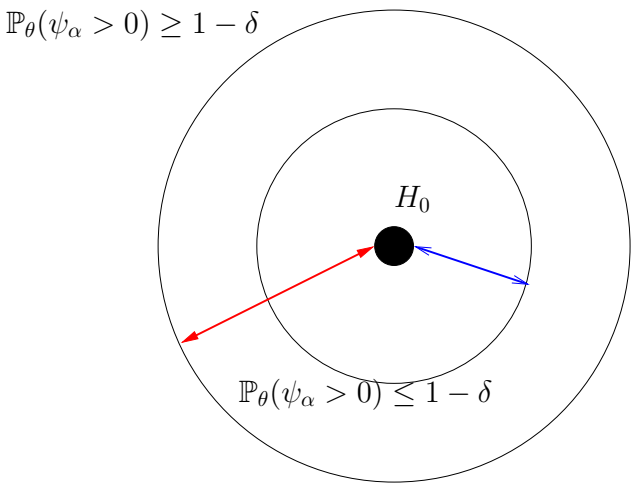


Minimax separation distance

$\mathbf{H_0} : \theta = 0$ against $\mathbf{H_1} : \theta \in \Theta[k, p] \setminus \{0\}$.
Fix $\delta > 0$. ψ_α test of Level α.

Separation distance of ψ_α :

$$\rho[\psi_\alpha, k, \mathbf{X}] := \inf \left\{ \rho > 0, \inf_{\theta \in \Theta[k,p],\ \|\mathbf{X}\theta\|_n \geq \sqrt{n}\rho\sigma} \mathbb{P}_{\theta,\sigma}[\psi_\alpha = 1] \geq 1 - \delta \right\} .$$



Minimax distance of separation

$$\rho^*[k, \mathbf{X}] := \inf_{\psi_\alpha} \rho[\psi_\alpha, k, \mathbf{X}] .$$

$$\rho^*[k] := \sup_{\mathbf{X}} \rho^*[k, \mathbf{X}]$$

Known variance σ^2

If the support of θ is known $\rightsquigarrow$ square of the parametric separation distance $\sqrt{k}/n$.

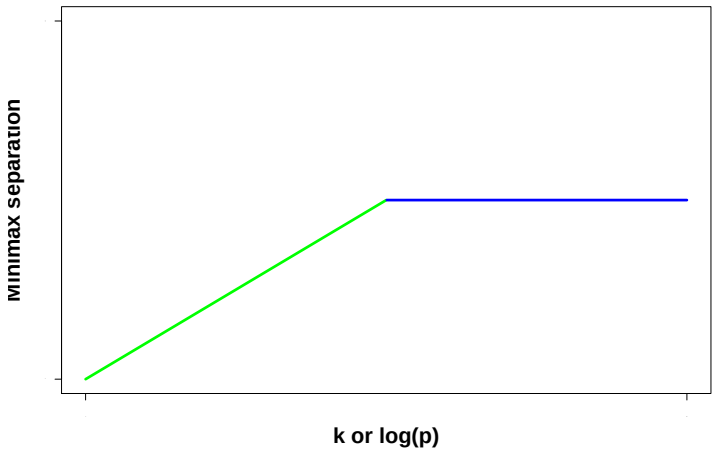
Known variance σ^2

If the support of θ is known $\rightsquigarrow$ square of the parametric separation distance $\sqrt{k}/n$.

Theorem

As long as $p \geq n \geq \square(\alpha, \delta)$ and $k \leq p^{1/3}$, we have

$$(\rho^*[k])^2 \simeq \square[\alpha, \delta] \left[\frac{k}{n} \log\left(\frac{ep}{k}\right) \wedge \frac{1}{\sqrt{n}} \right] .$$



Comments :

- If $k \log(ep/k)$ is small before $\sqrt{n}$, analogous to minimax risk for prediction. Analogous to Gaussian sequence model (*Baraud (02), Donoho/Jin (04)*).
- Large $(k, p) \Rightarrow$ parametric separation distance over $\mathbb{R}^n$.
- Adaptation to the sparsity is possible. (Bonferroni multiple testing procedure).

A near optimal procedure (Baraud 02)

$\mathcal{M}(k,p)$: subsets of $\{1,\dots,p\}$ of size k.
Π_m Orthogonal projection over $vect(\mathbf{X}_i,\ i \in m)$.

Reasonable dimension
For $m \in \mathcal{M}(k,p)$, $F_m := \|\Pi_m \mathbf{Y}\|_n^2/\sigma^2$.
Under H_0, $F_m \sim \chi^2(k)$.
Under H_1, $F_m \sim \chi^2(v^2,k)$ with $v^2 := \|\Pi_m \mathbf{X}\theta\|_n^2$.
$\rightsquigarrow$ if $\mathrm{supp}(\theta) = m$, $v^2 = \|\mathbf{X}\theta\|_n^2$.

A near optimal procedure (Baraud 02)

$\mathcal{M}(k,p)$: subsets of $\{1,\dots,p\}$ of size k.
Π_m Orthogonal projection over $vect(\mathbf{X}_i,\ i \in m)$.

Reasonable dimension
For $m \in \mathcal{M}(k,p)$, $F_m := \|\Pi_m \mathbf{Y}\|_n^2/\sigma^2$.
Under H_0, $F_m \sim \chi^2(k)$.
Under H_1, $F_m \sim \chi^2(v^2,k)$ with $v^2 := \|\Pi_m \mathbf{X}\theta\|_n^2$.
$\rightsquigarrow$ if $\mathrm{supp}(\theta) = m$, $v^2 = \|\mathbf{X}\theta\|_n^2$.

$\rightsquigarrow$ Bonferroni Procedure.

$$T_k := \sup_{m \in \mathcal{M}(k,p)} F_m - \bar{\chi}_k^{-1}\left(\alpha/|\mathcal{M}(k,p)|\right)$$

T_k is powerful if $\theta \in \Theta[k,p]$ et $\|\mathbf{X}\theta\|_n^2 \geq \square k \log(ep/k)\sigma^2$.

A near optimal procedure (Baraud 02)

$\mathcal{M}(k,p)$: subsets of $\{1,\dots,p\}$ of size k.
Π_m Orthogonal projection over $vect(\mathbf{X}_i,\ i \in m)$.

Reasonable dimension
For $m \in \mathcal{M}(k,p)$, $F_m := \|\Pi_m \mathbf{Y}\|_n^2/\sigma^2$.
Under H_0, $F_m \sim \chi^2(k)$.
Under H_1, $F_m \sim \chi^2(v^2,k)$ with $v^2 := \|\Pi_m \mathbf{X}\theta\|_n^2$.
$\rightsquigarrow$ if $\mathrm{supp}(\theta) = m$, $v^2 = \|\mathbf{X}\theta\|_n^2$.

$\rightsquigarrow$ Bonferroni Procedure.

$$T_k := \sup_{m \in \mathcal{M}(k,p)} F_m - \bar{\chi}_k^{-1}\left(\alpha/|\mathcal{M}(k,p)|\right)$$

T_k is powerful if $\theta \in \Theta[k,p]$ et $\|\mathbf{X}\theta\|_n^2 \geq \square k \log(ep/k)\sigma^2$.

Ultra-high dimension
If $k \geq k_*$, $T_n := \|\mathbf{Y}\|_n^2/\sigma^2 - \bar{\chi}_n^{-1}(\alpha)$.
T_n is powerful if $\|\mathbf{X}\theta\|_n^2 \geq \square\sqrt{n}\sigma^2$.

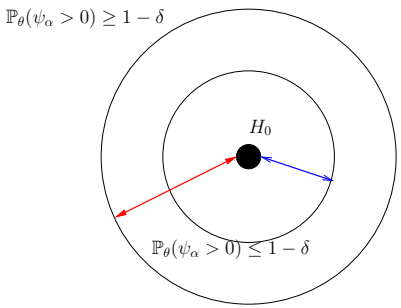
Unknown variance σ^2

ψ_α : $\sup_{\sigma>0} \mathbb{P}_{0,\sigma}[\psi_\alpha = 1] \leq \alpha$. Separation distance when the variance is unknown.

Unknown variance σ^2

ψ_α : $\sup_{\sigma>0} \mathbb{P}_{0,\sigma}[\psi_\alpha = 1] \leq \alpha$. Separation distance when the variance is unknown.

$$\rho_U[\psi_\alpha, k, \mathbf{X}] := \inf\left\{\rho > 0, \quad \inf_{\substack{\sigma>0,\ \theta\in\Theta[k,p],\\ \|\mathbf{X}\theta\|_n \geq \sqrt{n}\rho\sigma}} \mathbb{P}_{\theta,\sigma}[\psi_\alpha = 1] \geq 1-\delta\right\} .$$



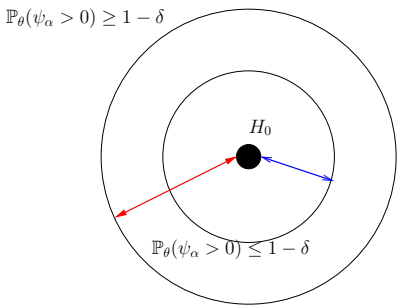
$$\rho^*_U[k, \mathbf{X}] := \inf_{\psi_\alpha} \rho_U[\psi_\alpha, k, \mathbf{X}] .$$

$$\rho^*_U[k] := \sup_{\mathbf{X}} \rho^*_U[k, \mathbf{X}]$$

Unknown variance σ^2

ψ_α : $\sup_{\sigma>0} \mathbb{P}_{0,\sigma}[\psi_\alpha = 1] \leq \alpha$. Separation distance when the variance is unknown.

$$\rho_U[\psi_\alpha, k, \mathbf{X}] := \inf \left\{ \rho > 0, \quad \inf_{\substack{\sigma>0,\ \theta\in\Theta[k,p], \\ \|\mathbf{X}\theta\|_n \geq \sqrt{n}\rho\sigma}} \mathbb{P}_{\theta,\sigma}[\psi_\alpha = 1] \geq 1-\delta \right\} .$$



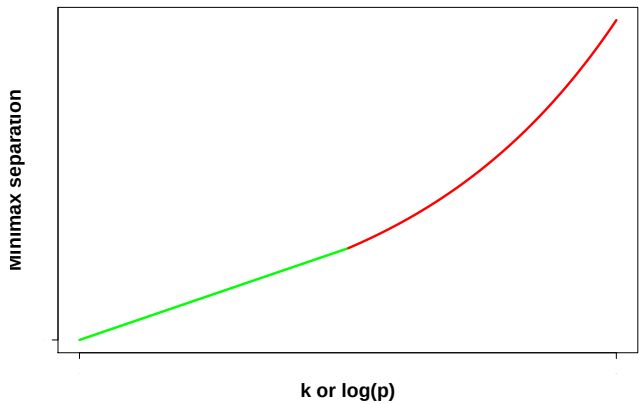
$$\rho^*_U[k, \mathbf{X}] := \inf_{\psi_\alpha} \rho_U[\psi_\alpha, k, \mathbf{X}] .$$

$$\rho^*_U[k] := \sup_{\mathbf{X}} \rho^*_U[k, \mathbf{X}]$$

Theorem

If $p \geq n \geq \square(\alpha, \delta)$ and $k \leq p^{1/3}$, we have

$$(\rho^*_U[k])^2 \simeq \square[\alpha,\delta] \frac{k}{n} \log\left(\frac{ep}{k}\right) \exp\left[\square[\alpha,\delta] \frac{k \log(ep/k)}{n}\right] .$$



Comments :

- If $k \log(ep/k) \leq \sqrt{n}$, same minimax separation distance as for known variance.
- Blow up in ultra-high dimension.
  Upper bound $\rightsquigarrow$ (*Baraud/Huet/Laurent (03)*) replace χ^2 tests by Fisher tests.

Main Idea (Le Cam)

Total variance distances

$$\|P - Q\|_{TV} = \sup_A |P(A) - Q(A)|$$

Simples Hypotheses :
We consider the hypothesis H_0 : $\theta = 0$ against H_1 $\theta = \theta_0$.
(α, β) Type I and II errors T.
$A = \mathbf{1}_{T=1}$ $\Rightarrow |P_0(A) - P_{\theta_0}(A)| = 1 - \beta - \alpha \leq \|P_0 - P_{\theta_0}\|_{TV}$.
$\rightsquigarrow \beta \geq 1 - \alpha - \|P_0 - P_{\theta_0}\|_{TV}$

Main Idea (Le Cam)

Total variance distances

$$\|P - Q\|_{TV} = \sup_A |P(A) - Q(A)|$$

Simples Hypotheses :
We consider the hypothesis H_0 : $\theta = 0$ against H_1 $\theta = \theta_0$.
(α, β) Type I and II errors T.
$A = \mathbf{1}_{T=1}$ $\Rightarrow |P_0(A) - P_{\theta_0}(A)| = 1 - \beta - \alpha \leq \|P_0 - P_{\theta_0}\|_{TV}$.
$\rightsquigarrow \beta \geq 1 - \alpha - \|P_0 - P_{\theta_0}\|_{TV}$

Composite hypotheses :
H_0 : $\theta = 0$ against H_1 : $\theta \in \{\theta_1, \theta_2, \ldots, \theta_r\}$
α type I error. Type II error is uniformly smaller β.
Idea : we consider a prior distribution μ on $\{\theta_1, \theta_2, \ldots, \theta_r\}$.
T test of P_0 against $P_\mu = \sum_{i=1}^r \mu_i P_{\theta_i}$.
$\rightsquigarrow \beta \geq 1 - \alpha - \|P_0 - P_\mu\|_{TV}$

Main Idea (Le Cam)

Total variance distances

$$\|P - Q\|_{TV} = \sup_A |P(A) - Q(A)|$$

Simples Hypotheses :
We consider the hypothesis H_0 : $\theta = 0$ against H_1 $\theta = \theta_0$.
(α, β) Type I and II errors T.
$A = \mathbf{1}_{T=1}$ $\Rightarrow |P_0(A) - P_{\theta_0}(A)| = 1 - \beta - \alpha \leq \|P_0 - P_{\theta_0}\|_{TV}$.
$\rightsquigarrow \beta \geq 1 - \alpha - \|P_0 - P_{\theta_0}\|_{TV}$

Composite hypotheses :
H_0 : $\theta = 0$ against H_1 : $\theta \in \{\theta_1, \theta_2, \ldots, \theta_r\}$
α type I error. Type II error is uniformly smaller β.
Idea : we consider a prior distribution μ on $\{\theta_1, \theta_2, \ldots, \theta_r\}$.
T test of P_0 against $P_\mu = \sum_{i=1}^r \mu_i P_{\theta_i}$.
$\rightsquigarrow \beta \geq 1 - \alpha - \|P_0 - P_\mu\|_{TV}$

Two ingredients :

1. A clever choice of μ
2. An upper bound of $\|P_0 - P_\mu\|_{TV}$.

Inverse problem

Loss function : $\|\theta - \widehat{\theta}\|_p^2/\sigma^2$.

$$\mathcal{RI}[k, \mathbf{X}] := \inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma}[\|\theta - \widehat{\theta}\|_p^2/\sigma^2] \ .$$

$\mathcal{RI}[k, \mathbf{X}]$ is in inversely proportional to the design.

Inverse problem

Loss function : $\|\theta - \widehat{\theta}\|_{\boldsymbol{p}}^2/\sigma^2$.

$$\mathcal{RI}[k, \mathbf{X}] := \inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,\boldsymbol{p}]} \mathbb{E}_{\theta,\sigma}[\|\theta - \widehat{\theta}\|_{\boldsymbol{p}}^2/\sigma^2] \ .$$

$\mathcal{RI}[k, \mathbf{X}]$ is in inversely proportional to the design.

$\rightsquigarrow$ Collection $\mathcal{D}_{\boldsymbol{n},\boldsymbol{p}}$ of designs $\mathbf{X}$ such that each column is normed to one.

$$\mathcal{RI}[k] := \inf_{\mathbf{X} \in \mathcal{D}_{\boldsymbol{n},\boldsymbol{p}}} \mathcal{RI}[k, \mathbf{X}] \ .$$

$\rightsquigarrow$ For the "best possible design", what is the minimax risk ?

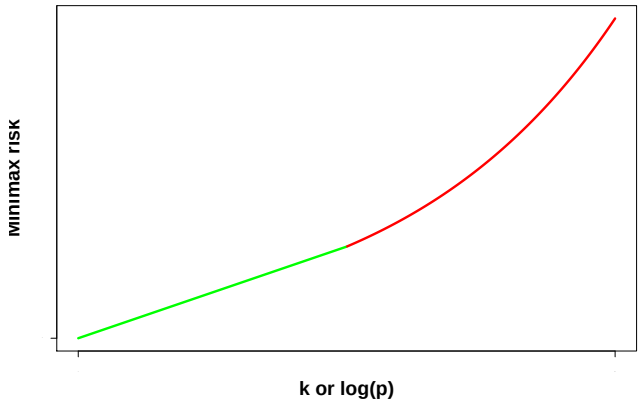
Theorem

Assume that $k \log(ep/k) \leq \square n$. *Then,*

$$\mathcal{RI}[k] \simeq \square k \log\left(\frac{ep}{k}\right) \ .$$

Assume that $k \log(ep/k) \gg n \log(n)$, *then*

$$\mathcal{RI}[k] \simeq \square \exp\left[\square \frac{k}{n} \log(p/k)\right]$$



Comments :

- In "reasonable" dimension, there exist designs such that the minimax risk is of the order $k \log\left(\frac{ep}{k}\right)$
  ex : Dantzig selector if $\mathbf{X}$ satisfies a restricted isometry property.
  Standard Gaussian design statisfies such a property.
- Blow up in ultra-high dimension. $\rightsquigarrow$ no design allows to recover θ.

Geometrical interpretation in ultra-high dimension

Proposition

For any design $\mathbf{X} \in \mathcal{D}_{n,p}$ *and any* $k \leq n \wedge p/2$, *we have*

$$\Phi_{2k,-}(\mathbf{X}) \leq Ck^2 \left(\frac{k}{p}\right)^{2k/n} \vee 1 \ .$$

Consider a sequence (k_n, p_n) *such that* $[k_n \log(p_n/k_n)]/\{n \log(n)\} \to \infty$.

$$\left(\frac{p_n}{k_n}\right)^{4k_n/n} \overset{\log}{\gtrsim} \mathcal{RI}[k_n] \overset{\log}{\sim} \inf_{\mathbf{X} \in D_{n,p_n}} \Phi^{-1}_{2k_n,-}(\mathbf{X}) \overset{\log}{\gtrsim} \left(\frac{p_n}{k_n}\right)^{2k_n/n} .$$

Support estimation and dimension reduction

Definition

The set $\mathcal{C}_k^p(\rho)$ is defined as $\theta \in \theta[k, p]$ such that θ contains exactly k non zero coefficient that are all equal to $\rho/\sqrt{k}$.

Support estimation and dimension reduction

Definition

The set $\mathcal{C}_k^p(\rho)$ is defined as $\theta \in \theta[k,p]$ such that θ contains exactly k non zero coefficient that are all equal to $\rho/\sqrt{k}$.

Hypothesis : $k \leq p^{1/3}$
X follows a standard Gaussian distribution.
$\sigma^2 = 1$.

Proposition (dimension reduction is almost impossible)

$\rho^2 = \square \frac{k}{n} \log\left(\frac{p}{k}\right) \exp\left[\square' \frac{k}{n} \log\left(\frac{p}{k}\right)\right]$.

There exists a constant $0 < \delta < 1$ such that for each set $\widehat{M}$ of $\{1,\dots,p\}$ of size $p_0 \leq p^\delta$, we have

$$\sup_{\theta \in \mathcal{C}_k^p(\rho)} \mathbb{P}_{\theta,1}\left[\mathrm{supp}(\theta) \not\subseteq \widehat{M}\right] \geq 1/8 \ .$$

Comments :

- In ultra-high dimension, it is almost impossible to estimate the support in θ.
- It is even almost impossible to reduce efficiently the dimension of the problem.

Simulations

$p = 5000$ and $p = 200$, $n = 50$.
$\sigma = 1$.
$\mathbf{X}$ follows a standard Gaussian distribution
$k = 1, \ldots, 15$.
$\theta_1 = \ldots = \theta_k = 4\sqrt{\log(p)/n} \approx 1.30$ (resp. 1.65) for $p = 200$ (resp. $p = 5000$) et
$\theta_{k+1} = \ldots = \theta_p = 0$.
We have $\|\theta\|^2 = 16k \log(p)/n$.

Simulations

$p = 5000$ and $p = 200$, $n = 50$.
$\sigma = 1$.
X follows a standard Gaussian distribution
$k = 1, \ldots, 15$.
$\theta_1 = \ldots = \theta_k = 4\sqrt{\log(p)/n} \approx 1.30$ (resp. 1.65) for $p = 200$ (resp. $p = 5000$) et
$\theta_{k+1} = \ldots = \theta_p = 0$.
We have $\|\theta\|^2 = 16k \log(p)/n$.

Dimension reduction procedure. We apply the SIS method (Lv and Fan) and the LASSO as a way to reduce dimension to a set $\widehat{M}^S$ of size $p_0 = 50$.
Power of the procedures :

$$\text{Puissance} := \frac{\text{Card}[\widehat{M}^S \cap \{1, \ldots, k\}]}{k} \ .$$

Simulations

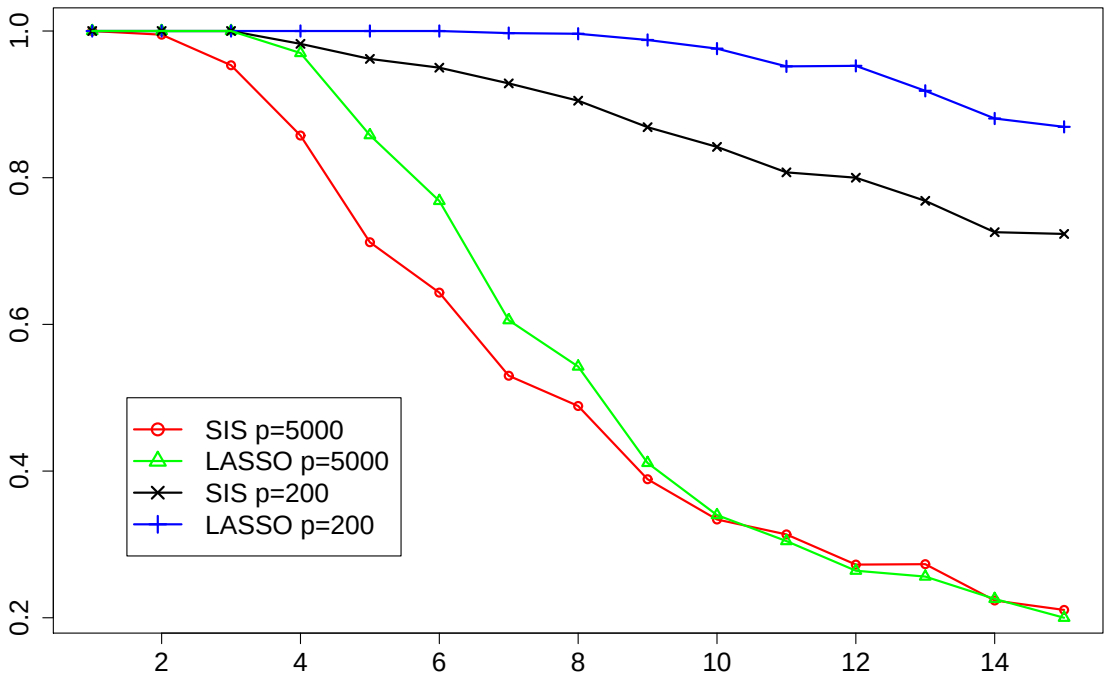


Figure: Power of the dimension reduction procedures (SIS and LASSO)

θ such that $\theta_1 = \ldots = \theta_k = u\sqrt{\log(p)/n}$ and $\theta_{k+1} = \ldots = \theta_p = 0$.
Compute u_k^* the smallest u such that $\widehat{M}^L$ has a power larger than 0.9.

$\rightsquigarrow$ u_k^* corresponds to the minimal intensity of the signal so that the reduction procedures forgets non relevant covariates.

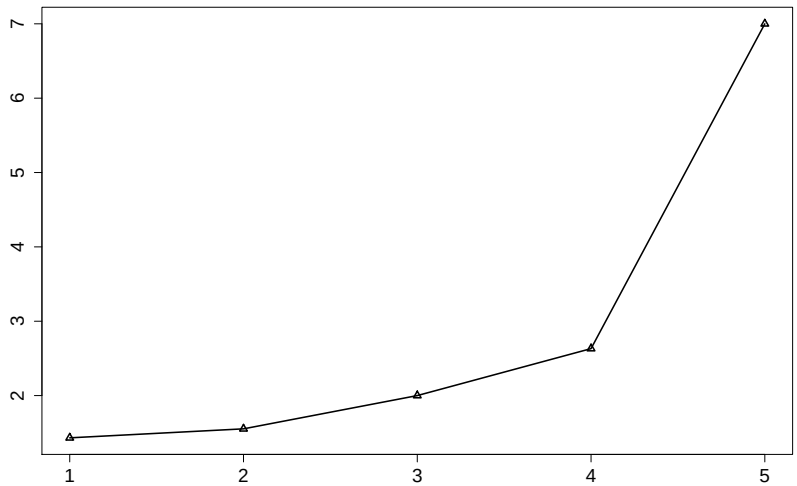


Figure: Minimal signal u_k^* as a function of k.

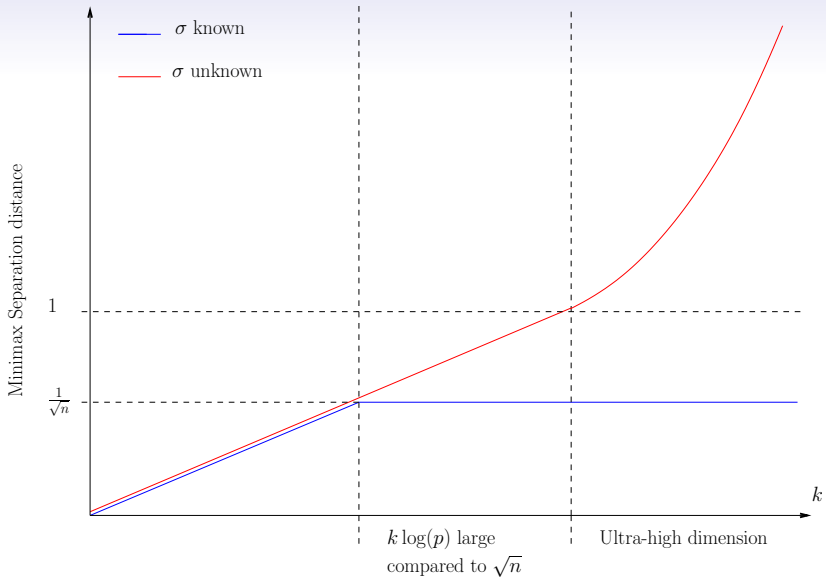


Figure: Tests ($\mathbf{P_2}$)

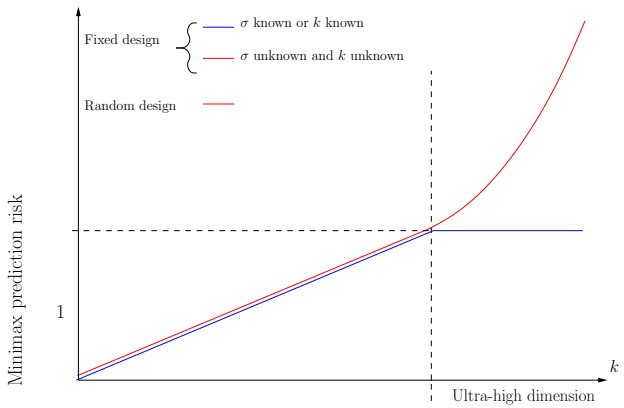


Figure: Prediction ($\mathbf{P_1}$)

~

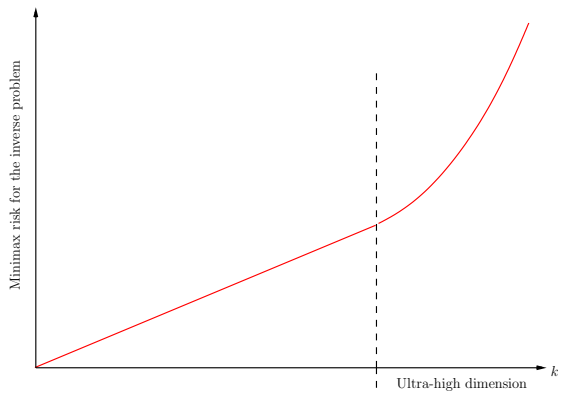


Figure: Inverse ($\mathbf{P_3}$)

Conclusion

Transition when $k \log(p/k)/n$ is large.

Conclusion

Transition when $k \log(p/k)/n$ is large.

Several phenomenons are occurring in ultra-high dimension

1. Minimax risks are blowing up.
2. The knowledge of the variance is crucial for some problems.
3. Difference between the fixed design and random design.
4. The problems related to inverse problems become too difficult.

Conclusion

What does the sentence "$k \log(p/k)/n$ is large" mean ? One heuristic : 1/2

Rule of thumb :
$p = 5000$ genes and $n = 40$ microarray experiments : $\rightsquigarrow 4 \log(p/4)/n \simeq 0.57$

Conclusion

What does the sentence "$k \log(p/k)/n$ is large" mean ? One heuristic : $1/2$

Rule of thumb :
$p = 5000$ genes and $n = 40$ microarray experiments : $\rightsquigarrow 4 \log(p/4)/n \simeq 0.57$

What can we do for ultra-high dimensional models ?

1. Take n larger (sometimes expensive)
2. Take p smaller $\rightsquigarrow$ Okay, if some covariates by prior knowledge.
   $\rightsquigarrow$ No ! if dimension reduction are applied.
3. Use some prior knowledge. $\rightsquigarrow$ okay, if the prior probability of each model is larger than e^{-n} !

Raskutti, Wainwright, and Yu (2010). Minimax rates of estimations for high-dimensional regression over l_q balls. Preprint UC Berkely.

Rigollet and Tsybakov (2010). Exponential Screening and optimal rates of sparse estimation. `http://arxiv.org/pdf/1003.2654`

V. (2010). Minimax risks for sparse regressions : Ultra-high-dimensional phenomenons. `http://arxiv.org/abs/1008.0526v2`