Model Selection in High Dimension

Jalal Fadili
Normandie Université-ENSICAEN, GREYC

Mathematical coffees 2017

Model selection: what for ?

- A key conceptual tool for **dimension reduction** and exploiting hidden structures in **high-dimensional** data.

Model selection: what for ?

- A key conceptual tool for **dimension reduction** and exploiting hidden structures in **high-dimensional** data.
- The gist:
  - **Compare** different statistical models each for a possible hidden structure.
  - **Select** the one that is more suited for your task.

Model selection: what for ?

- A key conceptual tool for **dimension reduction** and exploiting hidden structures in **high-dimensional** data.
- The gist:
  - **Compare** different statistical models each for a possible hidden structure.
  - **Select** the one that is more suited for your task.
- Challenges and questions:
  - Data-driven model selection, i.e. without an oracle.
  - Guarantees ? In terms of what ?
  - Optimality.
  - Computational issues.
  - Tractable procedures (recall MC on sparsity and CS).

Gaussian regression

$$y_i = f^\star(x_i) + \varepsilon_i$$

- $x_1, \ldots, x_n$: design vectors in $\mathbb{R}^p$.
- $f^\star : \mathbb{R}^p \to \mathbb{R}$: unknown regression function.
- $(\varepsilon_1, \ldots, \varepsilon_n)$ are independent and identically distributed $\mathcal{N}(0, \sigma^2)$.

Gaussian regression

$$y_i = f^\star(x_i) + \varepsilon_i$$

- $x_1, \ldots, x_n$: design vectors in $\mathbb{R}^p$.
- $f^\star : \mathbb{R}^p \to \mathbb{R}$: unknown regression function.
- $(\varepsilon_1, \ldots, \varepsilon_n)$ are independent and identically distributed $\mathcal{N}(0, \sigma^2)$.

Goal

Estimate $f^\star$ from data $\{(x_i, y_i)\}_{1 \le i \le n}$

Gaussian regression

$$y_i = f^\star(x_i) + \varepsilon_i$$

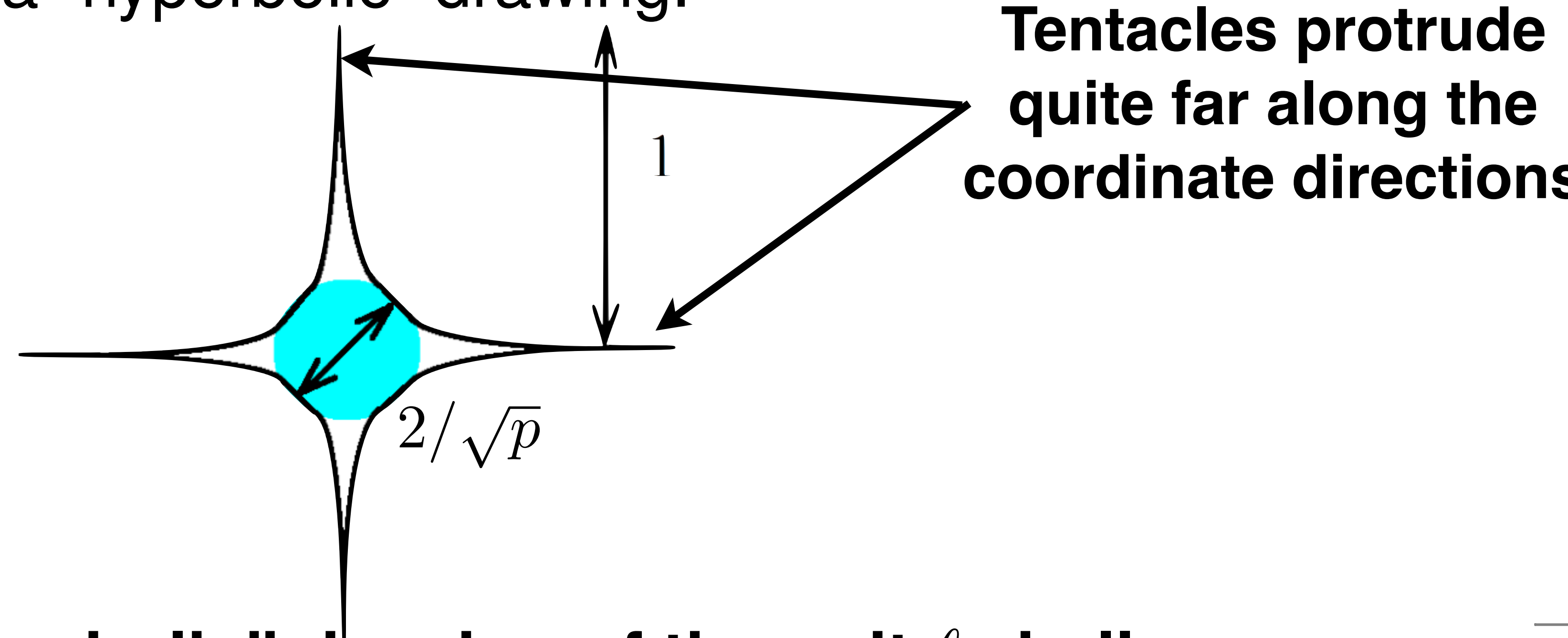
- $x_1, \ldots, x_n$: design vectors in $\mathbb{R}^p$.
- $f^\star : \mathbb{R}^p \to \mathbb{R}$: unknown regression function.
- $(\varepsilon_1, \ldots, \varepsilon_n)$ are independent and identically distributed $\mathcal{N}(0, \sigma^2)$.
- Curse of dimensionality :
  - $f^\star$ Lipschitz.
  - Need $n \asymp (1/\sigma)^{p+2}$ for root mean-square error σ.
- Blessings of dimensionality :
  - Concentration of measure.
  - High-dimensional geometry of convex bodies.
  - Asymptotic regime.
  - Approach to continuum.

Convex bodies in high-dimension

- A convex body C usually consists of a bulk and outliers.
  - The bulk makes up most of the volume of C, but it is usually small in diameter.
  - The outliers contribute little to the volume, but they are large in diameter.
- If C is properly scaled, the bulk usually looks like a Euclidean ball.
- The outliers look like thin, long tentacles.
- The volume in high dimensions scales differently than in low dimensions :
  - dilate by 2 increases volume by 2^p.
  - the tentacles contain exponentially less volume than the bulk.
- This is better captured in a ”hyperbolic” drawing.

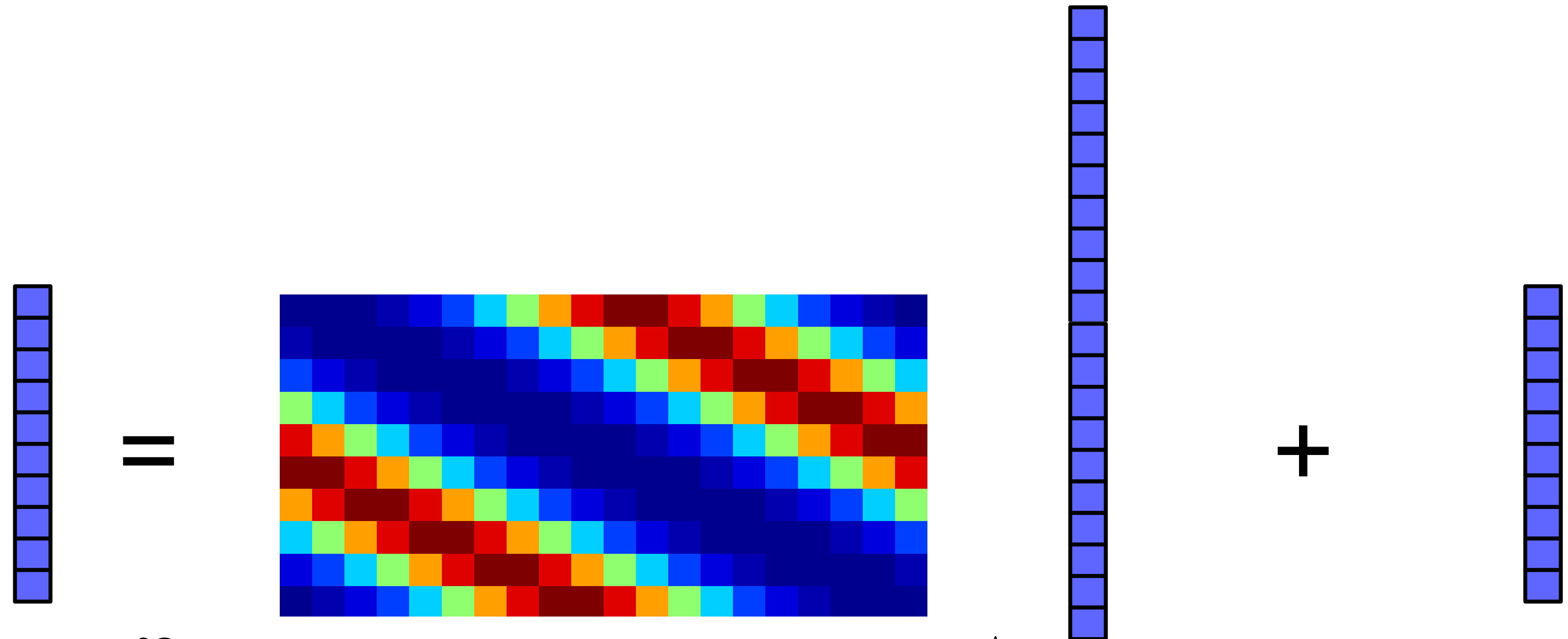
Convex bodies in high-dimension

- A convex body C usually consists of a bulk and outliers.
  - The bulk makes up most of the volume of C, but it is usually small in diameter.
  - The outliers contribute little to the volume, but they are large in diameter.
- If C is properly scaled, the bulk usually looks like a Euclidean ball.
- The outliers look like thin, long tentacles.
- The volume in high dimensions scales differently than in low dimensions :
  - dilate by 2 increases volume by 2^p.
  - the tentacles contain exponentially less volume than the bulk.
- This is better captured in a ”hyperbolic” drawing.



Milman’s “hyperbolic” drawing of the unit ℓ_1-ball

Blessings in action

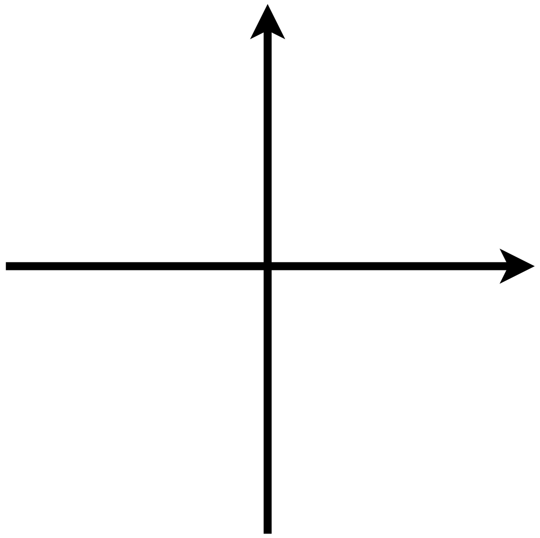


$y \in \mathbb{R}^n \quad X \in \mathbb{R}^{n \times p} \quad \beta^\star \in \mathbb{R}^p \quad \varepsilon \in \mathbb{R}^n$

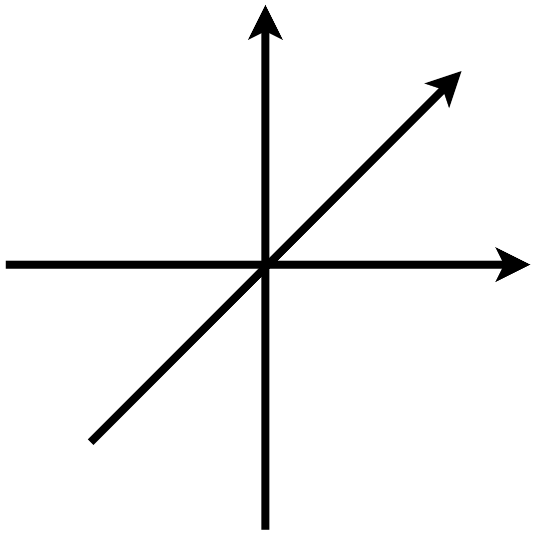
Estimate $\beta^\star$ from data (X, y)

- X is a fat design matrix :
  - **High-dimensional** setting : tendency to large p.
- $\beta^\star$ has some **intrinsically small-dimensional** structure :
  - Presumably a few key attributes.
  - Unwilling to specify in advance.
- $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$.

Blessings in action: Sparsity



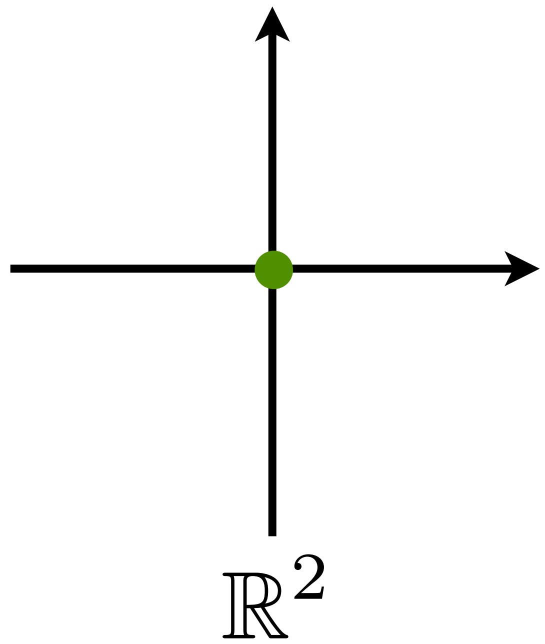
$\mathbb{R}^2$

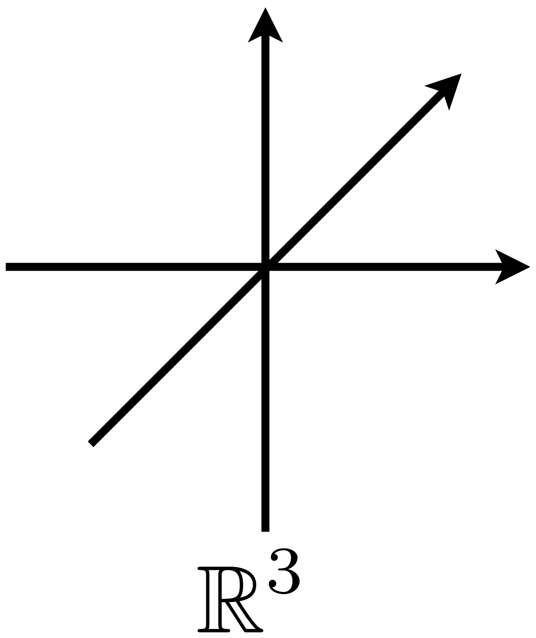


$\mathbb{R}^3$

Blessings in action: Sparsity

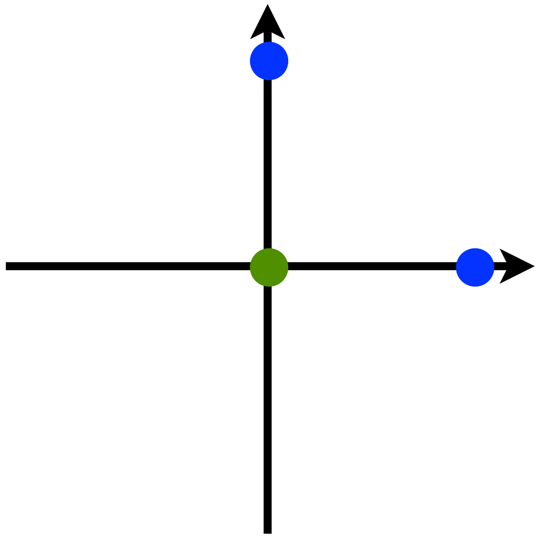
- 0-sparse



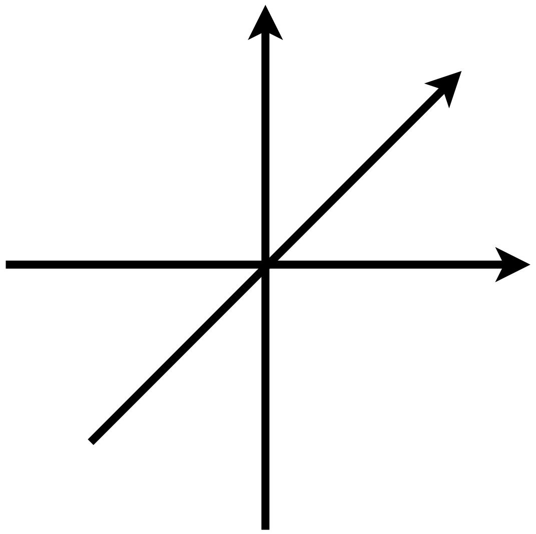


Blessings in action: Sparsity

- 0-sparse
- 1-sparse



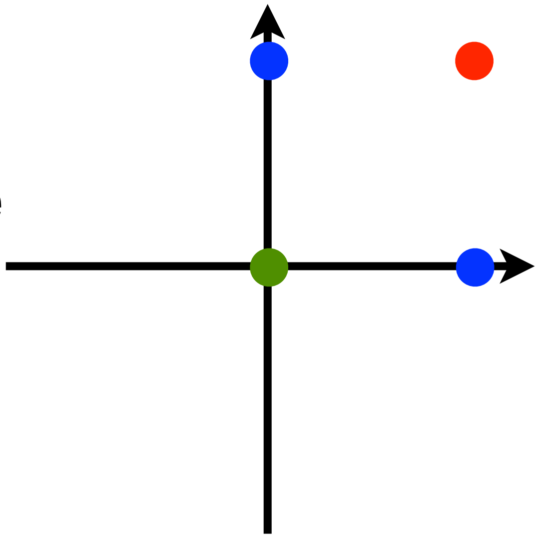
$\mathbb{R}^2$



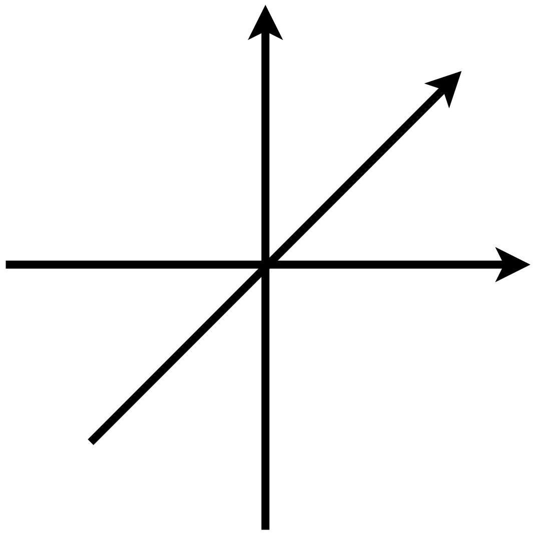
$\mathbb{R}^3$

Blessings in action: Sparsity

- 0-sparse
- 1-sparse
- 2-sparse = dense



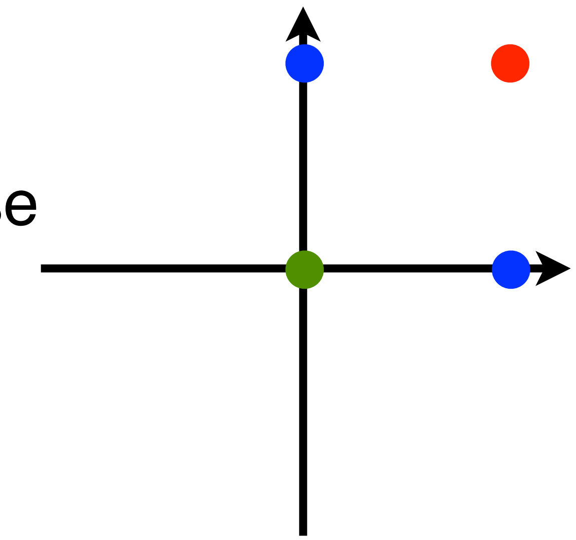
$\mathbb{R}^2$



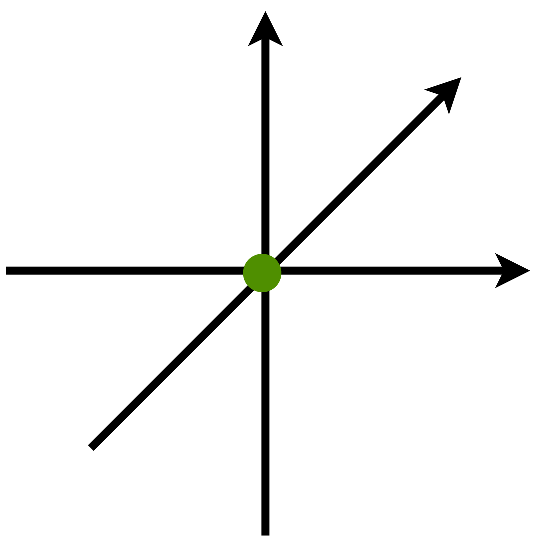
$\mathbb{R}^3$

Blessings in action: Sparsity

- 0-sparse
- 1-sparse
- 2-sparse = dense



$\mathbb{R}^2$

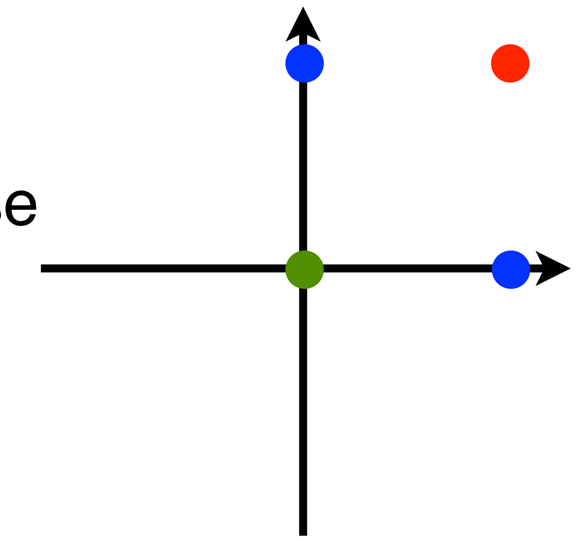


$\mathbb{R}^3$

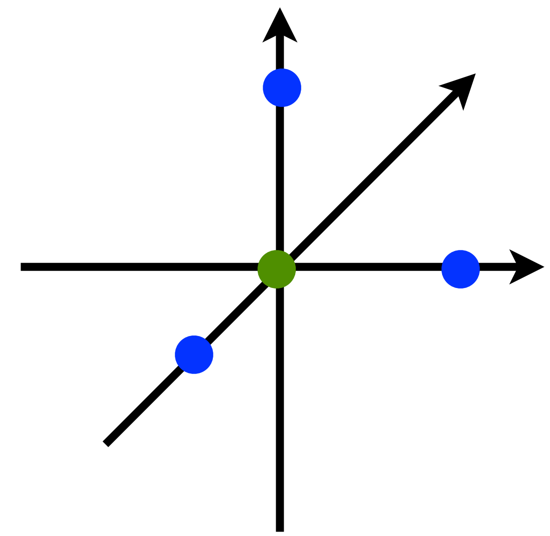
- 0-sparse

Blessings in action: Sparsity

- 0-sparse
- 1-sparse
- 2-sparse = dense



$\mathbb{R}^2$

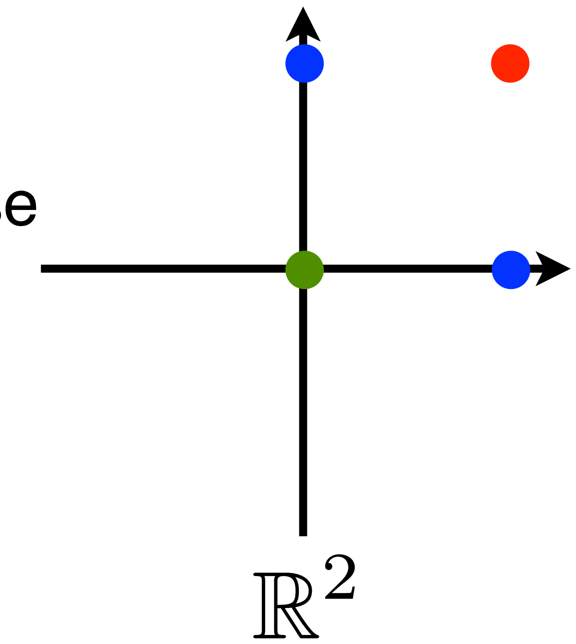


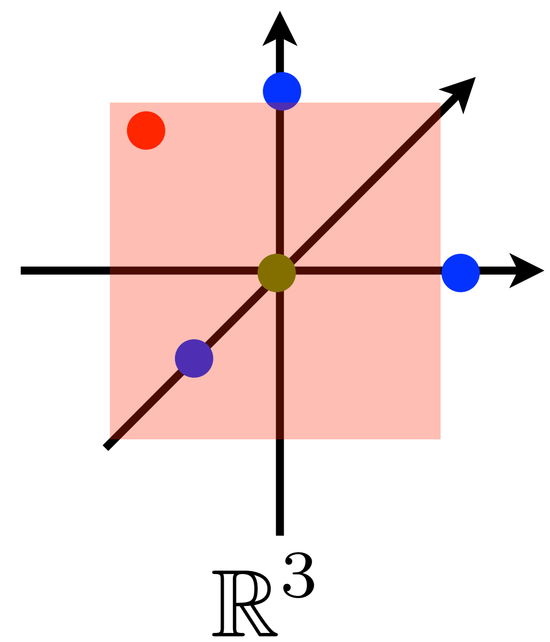
$\mathbb{R}^3$

- 0-sparse
- 1-sparse

Blessings in action: Sparsity

- 0-sparse
- 1-sparse
- 2-sparse = dense

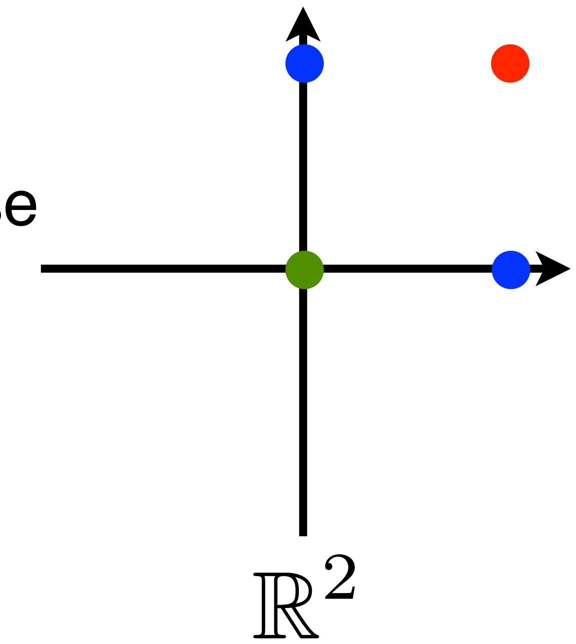


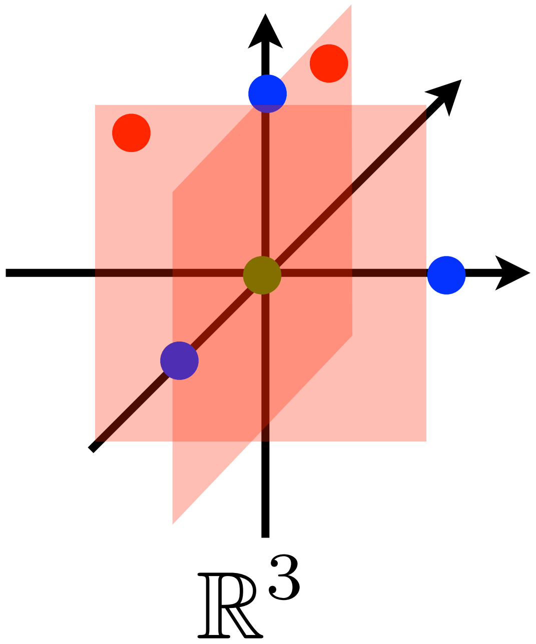


- 0-sparse
- 1-sparse
- 2-sparse

Blessings in action: Sparsity

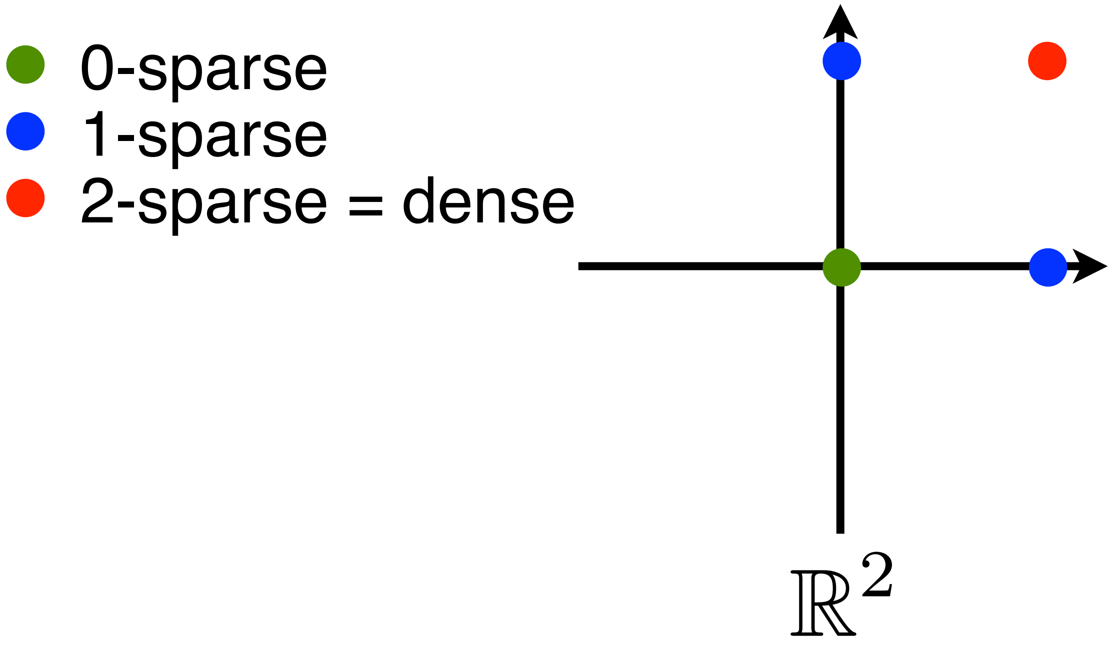
- 0-sparse
- 1-sparse
- 2-sparse = dense

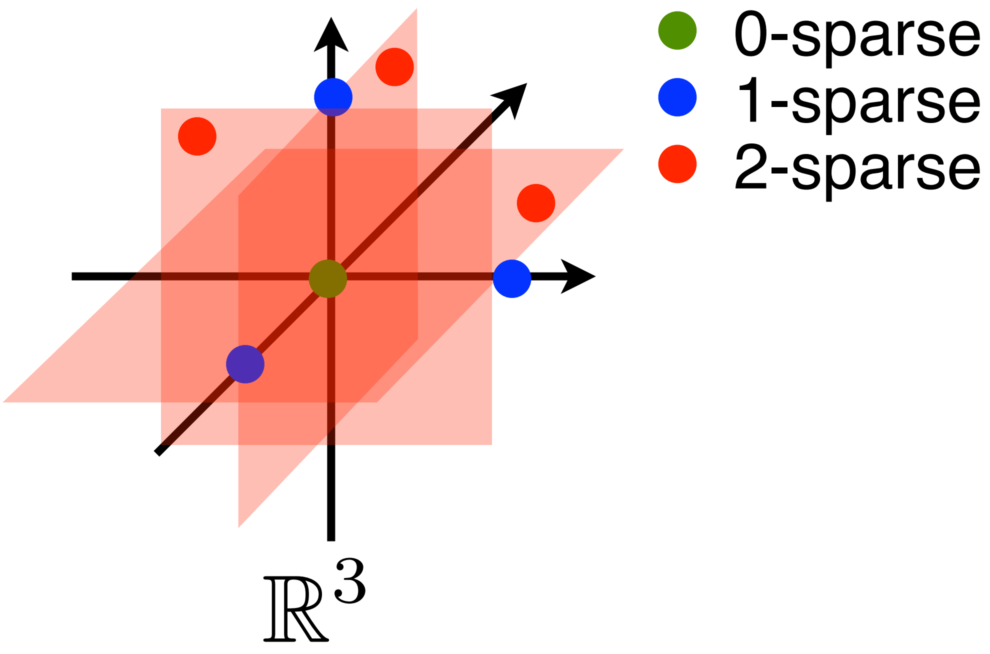




- 0-sparse
- 1-sparse
- 2-sparse

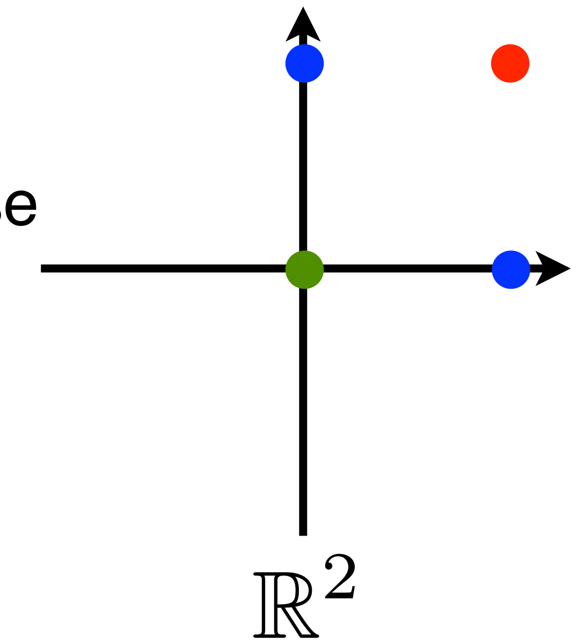
Blessings in action: Sparsity

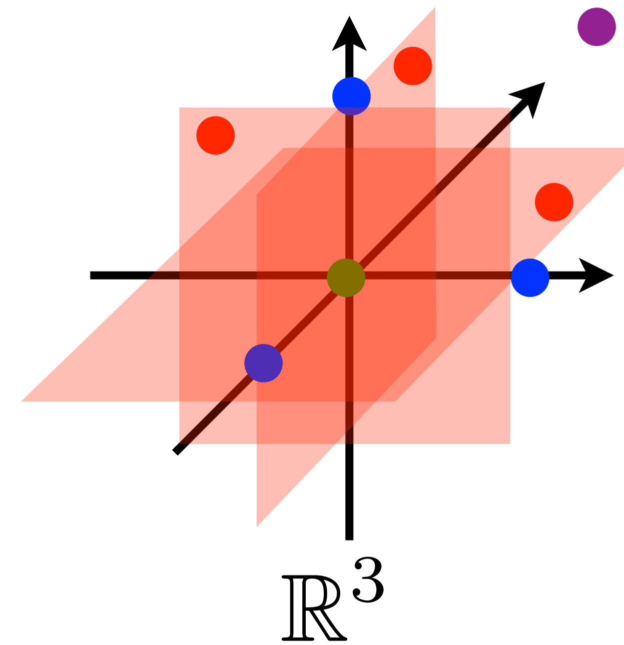




Blessings in action: Sparsity

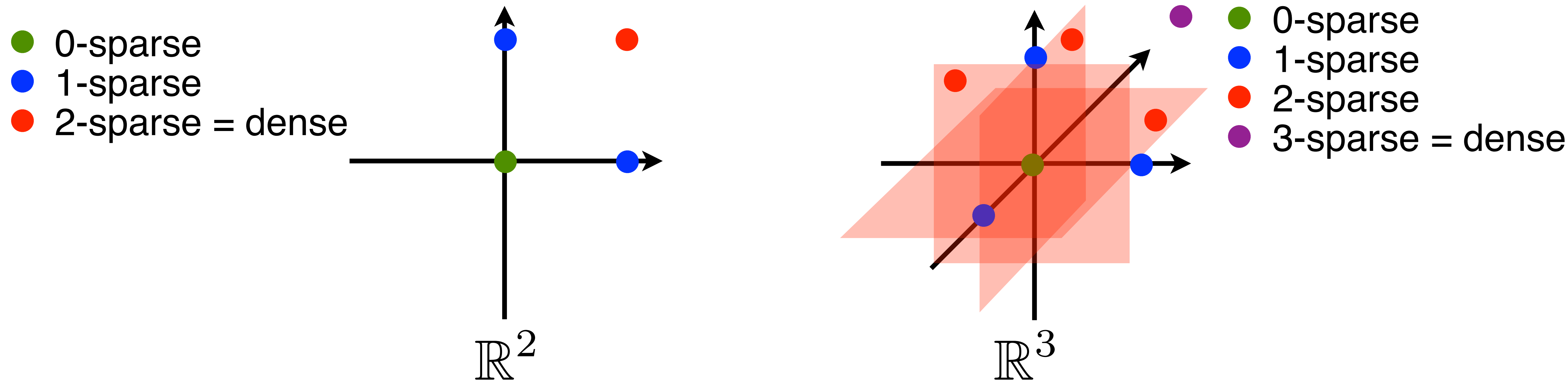
- 0-sparse
- 1-sparse
- 2-sparse = dense





- 0-sparse
- 1-sparse
- 2-sparse
- 3-sparse = dense

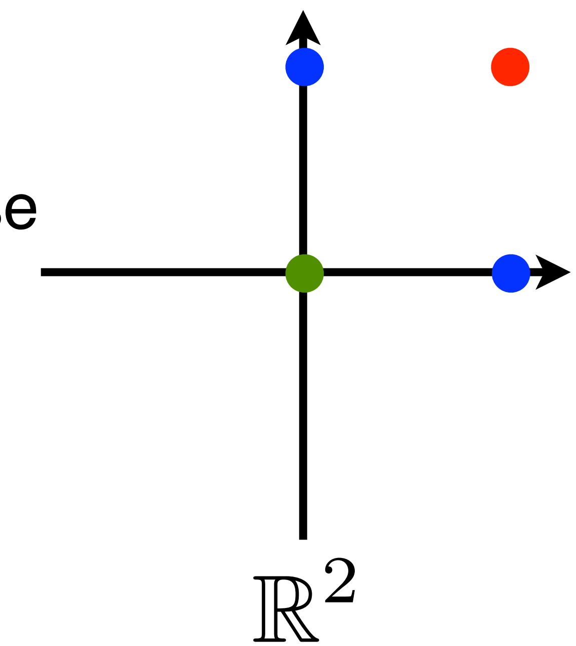
Blessings in action: Sparsity

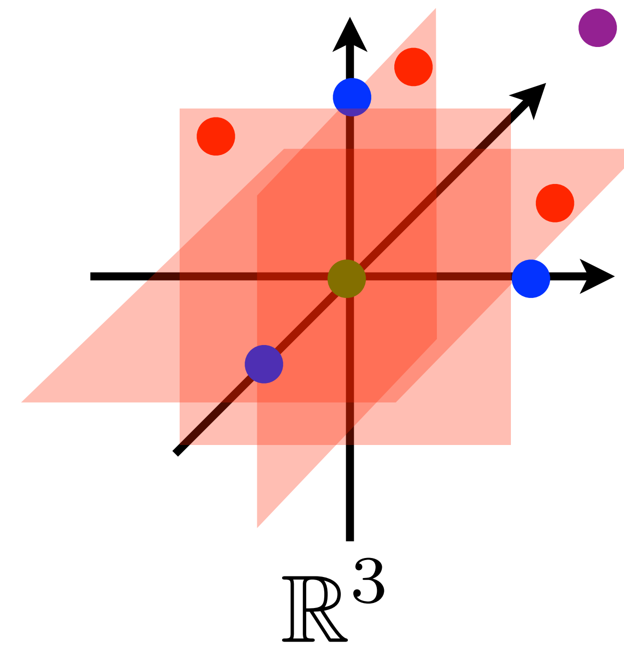


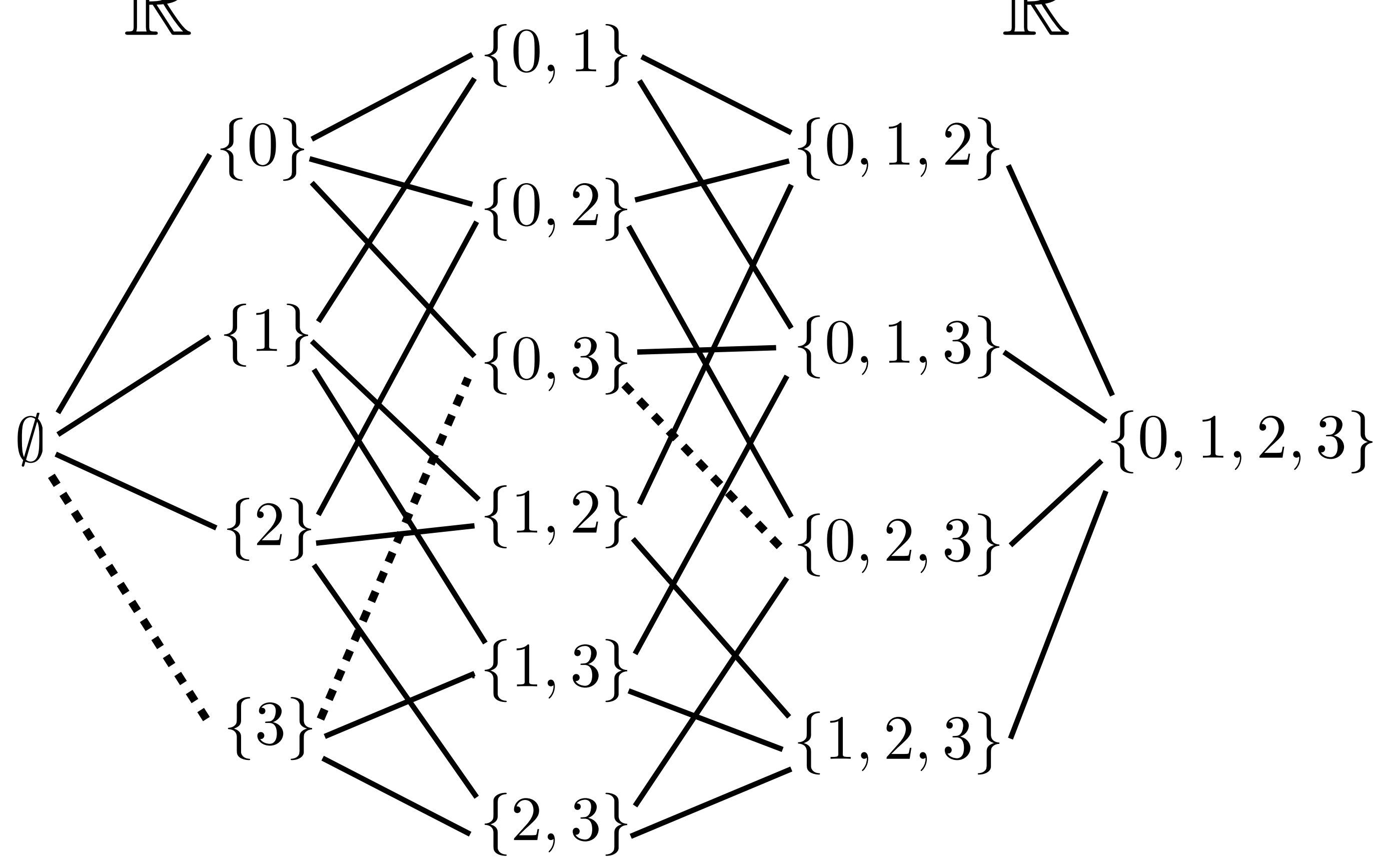
$$\mathrm{supp}(\beta) = \{i = 1, \cdots, n : \ \beta_i \neq 0\}$$
$$\|\beta\|_0 = \#\,\mathrm{supp}(\beta)$$

(Not a norm : not positively homogenenous)

Blessings in action: Sparsity







$$\mathscr{I} = \bigcup_{\beta \in \mathbb{R}^p} \mathrm{supp}(\beta)$$

$$|\mathscr{I}| = 2^p$$

Model of s-sparse vectors : a union of subspaces

$$\mathcal{M}_s = \bigcup_i \{V_i = \mathrm{span}\left((e_j)_{1 \le j \le n}\right) : \ \dim(V_i) = s\}.$$

Models and oracle

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon \qquad I^\star = \mathrm{supp}(\beta^\star)$$

$$\mathscr{I} = \bigcup_{\beta \in \mathbb{R}^p} \mathrm{supp}(\beta)$$

- Known support $I^\star$:

$$\widehat{f} \in X_{I^\star} \operatorname*{Argmin}_{\beta \in \mathbb{R}^{|I^\star|}} \frac{1}{2\sigma^2} \|y - X_{I^\star}\beta\|_2^2 + \frac{n}{2} \log(2\pi\sigma^2)$$

$$\iff \widehat{f} = \mathrm{Proj}_{V_{I^\star}}(y), \;\; V_{I^\star} = \mathrm{Span}(X_{I^\star}).$$

Models and oracle

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon \qquad\qquad I^\star = \mathrm{supp}(\beta^\star)$$

$$\mathscr{I} = \bigcup_{\beta \in \mathbb{R}^p} \mathrm{supp}(\beta)$$

- Known support $I^\star$:

$$\widehat{f} \in X_{I^\star} \underset{\beta \in \mathbb{R}^{|I^\star|}}{\mathrm{Argmin}} \frac{1}{2\sigma^2} \|y - X_{I^\star}\beta\|_2^2 + \frac{n}{2}\log(2\pi\sigma^2)$$

$$\iff \widehat{f} = \mathrm{Proj}_{V_{I^\star}}(y), \quad V_{I^\star} = \mathrm{Span}(X_{I^\star}).$$

- Unknown support $I^\star$ in practice :
  - Consider a collection $\{V_I, I \in \mathscr{I}\}$ of linear subspaces $\mathbb{R}^n$, called **models** ;
  - Associate to each subspace V_I the constrained maximum likelihood estimators $\widehat{f}_I = \mathrm{Proj}_{V_I}(y)$;
  - Estimate $f^\star$ by the **best** estimator among the collection $\{\widehat{f}_I, I \in \mathscr{I}\}$.
- Meaning of **best** ?

Models and oracle

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon \qquad I^\star = \mathrm{supp}(\beta^\star)$$

$$\mathscr{I} = \bigcup_{\beta \in \mathbb{R}^p} \mathrm{supp}(\beta)$$

- Risk to measure quality of an estimator :

$$R(\widehat{f}) = \mathbb{E}\left[\|\widehat{f} - f^\star\|_2^2\right].$$

- The best estimator in terms of the so-called **oracle** estimator :

$$\widehat{f}_{I_b} = \mathrm{Proj}_{V_{I_b}}(y), \quad I_b \in \underset{I \in \mathscr{I}}{\mathrm{Argmin}}\, R(\widehat{f}_I).$$

$$R(\widehat{f}_I) = \underbrace{\|\mathrm{Proj}_{V_I^\perp}(f^\star)\|_2^2}_{\text{Bias}} + \underbrace{\dim(V_I)\sigma^2}_{\text{Variance}}.$$

- The oracle model V_{I_b} is that in the collection $\{V_I : I \in \mathscr{I}\}$ which achieves the best bias-variance trade-off :
  - Bias decreases with dimension of V_I.
  - Variance increases with dimension of V_I.

Unbiased risk estimator

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon \qquad I^\star = \mathrm{supp}(\beta^\star)$$

$$R(\widehat{f}_I) = \underbrace{\|\mathrm{Proj}_{V_I^\perp}(f^\star)\|_2^2}_{\text{Bias}} + \underbrace{\dim(V_I)\sigma^2}_{\text{Variance}}. \qquad \widehat{f}_I = \mathrm{Proj}_{V_I}\, y$$

- Unfortunately, $R(\widehat{f}_I)$ cannot be computed in practice : depends on $f^\star$ which is unknown to the user.
- Replace $R(\widehat{f}_I)$ by a **good**, yet **computable**, estimate.
- Observe that

$$\begin{aligned}
\mathbb{E}\left[\|y - \widehat{f}_I\|_2^2\right] &= \mathbb{E}\left[\|\mathrm{Proj}_{V_I^\perp}(f^\star + \varepsilon)\|_2^2\right] \\
&= \mathbb{E}\left[\|\mathrm{Proj}_{V_I^\perp} f^\star\|_2^2\right] + \mathbb{E}\left[\left\langle \mathrm{Proj}_{V_I^\perp} f^\star, \varepsilon\right\rangle\right] + \mathbb{E}\left[\|\mathrm{Proj}_{V_I^\perp} \varepsilon\|_2^2\right] \\
&= \mathbb{E}\left[\|\mathrm{Proj}_{V_I^\perp} f^\star\|_2^2\right] + (n - \dim(V_I))\sigma^2 = R(\widehat{f}_I) + (n - 2\dim(V_I))\sigma^2.
\end{aligned}$$

- Unbiased risk estimator :

$$\mathbb{E}\left[\widehat{R}(\widehat{f}_I)\right] = R(\widehat{f}_I), \quad \widehat{R}(\widehat{f}_I) = \|y - \widehat{f}_I\|_2^2 + (2\dim(V_I) - n)\sigma^2.$$

Unbiased risk estimator

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon \qquad\qquad I^\star = \mathrm{supp}(\beta^\star)$$

$$R(\widehat{f}_I) = \overset{\text{Bias}}{\boxed{\left\| \mathrm{Proj}_{V_I^\perp}(f^\star) \right\|_2^2}} + \overset{\text{Variance}}{\boxed{\dim(V_I)\sigma^2}}. \qquad\qquad \widehat{f}_I = \mathrm{Proj}_{V_I} y$$

- Unfortunately, $R(\widehat{f}_I)$ cannot be computed in practice : depends on $f^\star$ which is unknown to the user.
- Replace $R(\widehat{f}_I)$ by a **good**, yet **computable**, estimate.
- Observe that

$$\begin{aligned}
\mathbb{E}\left[\|y - \widehat{f}_I\|_2^2\right] &= \mathbb{E}\left[\left\| \mathrm{Proj}_{V_I^\perp}(f^\star + \varepsilon)\right\|_2^2\right] \\
&= \mathbb{E}\left[\left\| \mathrm{Proj}_{V_I^\perp} f^\star\right\|_2^2\right] + \mathbb{E}\left[\left\langle \mathrm{Proj}_{V_I^\perp} f^\star, \varepsilon \right\rangle\right] + \mathbb{E}\left[\left\| \mathrm{Proj}_{V_I^\perp} \varepsilon\right\|_2^2\right] \\
&= \mathbb{E}\left[\left\| \mathrm{Proj}_{V_I^\perp} f^\star\right\|_2^2\right] + (n - \dim(V_I))\sigma^2 = R(\widehat{f}_I) + (n - 2\dim(V_I))\sigma^2.
\end{aligned}$$

- Unbiased risk estimator :

$$\mathbb{E}\left[\widehat{R}(\widehat{f}_I)\right] = R(\widehat{f}_I), \quad \widehat{R}(\widehat{f}_I) = \|y - \widehat{f}_I\|_2^2 + (2\dim(V_I) \overset{\text{AIC}}{\cancel{-n}})\sigma^2.$$

Akaike Information Criterion (AIC)

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon \qquad\qquad I^\star = \mathrm{supp}(\beta^\star)$$

$$\mathscr{I} = \bigcup_{\beta \in \mathbb{R}^p} \mathrm{supp}(\beta)$$

$$\mathrm{AIC}(\widehat{f}_I) = \left\| \mathrm{Proj}_{V_I^\perp} y \right\|_2^2 + 2\dim(V_I)\sigma^2.$$

- Select the model that minimizes AIC :

$$I_{\mathrm{AIC}} \in \operatorname*{Argmin}_{I \in \mathscr{I}} \mathrm{AIC}(\widehat{f}_I)$$

- Popular and simple to implement, but ... yields very poor results in high dimension (number of models grows fast with dimension p as in the sparse case).
- Simple justification assuming $X = \mathrm{Id}$:
  - $\mathrm{AIC}(\widehat{f}_I) = \|y\|_2^2 + \sum_{i \in I}(2\sigma^2 - |y_i|^2)$.
  - $I_{\mathrm{AIC}} = \left\{ i : |y_i|^2 > 2\sigma^2 \right\}$ (hard thresholding).
  - If $f^\star = 0$: $|I_{\mathrm{AIC}}| \sim \mathcal{B}(p, q)$, $q = \Pr(|Z| > \sqrt{2})$, $Z \sim \mathcal{N}(0,1)$.
  - $\mathbb{E}\left[|I_{\mathrm{AIC}}| \right] = pq \approx 0.157p$ while $I_{\mathrm{oracle}} = \emptyset$! ! !

Bayesian Information Criterion (BIC)

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon \qquad\qquad I^\star = \mathrm{supp}(\beta^\star)$$

$$\mathscr{I} = \bigcup_{\beta \in \mathbb{R}^p} \mathrm{supp}(\beta)$$

- Bayesian paradigm, $I^\star$, $f^\star$ and ε are random :
  - $I^\star$ sampled from $\{\pi_I : I \in \mathscr{I}\}$;
  - $f^\star$ sampled from $dP(f|I^\star)$ on $V_{I^\star}$;
  - Generate y.
- $P(I|y) \propto \int_{f \in V_I} \pi_I \exp(-\|y - f\|_2^2/(2\sigma^2))dP(f|I)$.
- Posterior likelihood ratio :

$$\log\left(\frac{P(I|y)}{P(I'|y)}\right) \sim_{n \to +\infty} \frac{\|y - \widehat{f}_{I'}\|_2^2 - \|y - \widehat{f}_I\|_2^2}{2\sigma^2} + \log(n)\frac{\dim(V_{I'}) - \dim(V_I)}{2}$$
$$+ \log(\pi_I/\pi_{I'}) + O(1).$$

- The most likely model would minimize :

$$\|y - \widehat{f}_I\|_2^2 + \sigma^2 \dim(V_I)\log(n) + 2\sigma^2 \log(\pi_I^{-1}).$$

- If the prior π_I is uniform, we get the BIC :

$$I_{\mathrm{BIC}} \in \operatorname*{Argmin}_{I \in \mathscr{I}} \mathrm{BIC}(I) = \|y - \widehat{f}_I\|_2^2 + \sigma^2 \dim(V_I)\log(n).$$

Bayesian Information Criterion (BIC)

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon \qquad I^\star = \mathrm{supp}(\beta^\star)$$

$$\mathscr{I} = \bigcup_{\beta \in \mathbb{R}^p} \mathrm{supp}(\beta)$$

$$I_{\mathrm{BIC}} \in \operatorname*{Argmin}_{I \in \mathscr{I}} \mathrm{BIC}(I) = \left\| y - \widehat{f}_I \right\|_2^2 + \sigma^2 \dim(V_I) \log(n).$$

- Again, popular and simple poor results in high dimension.
- Simple justification assuming $X = \mathrm{Id}$:
  - $I_{\mathrm{BIC}} = \left\{ i : |y_i|^2 > \log(p)\sigma^2 \right\}$ (hard thresholding).
  - If $f^\star = 0$: $\mathbb{E}\left[|I_{\mathrm{BIC}}| \right] = pq \asymp \sqrt{\frac{2p}{\pi \log(p)}}$, since

    $$q = \Pr(|Z| > \sqrt{\log(p)}) \asymp \sqrt{\frac{2}{\pi \log(p)}} e^{-\log(p)/2} = \sqrt{\frac{2}{\pi p \log(p)}}.$$

    $Z \sim \mathcal{N}(0, 1)$.
  - Again $\mathbb{E}\left[|I_{\mathrm{BIC}}| \right]$ grows with p while $I_{\mathrm{oracle}} = \emptyset$! ! !

Penalized empirical risk minimization

- How to avoid the selection of a model V_I with a large dimension ?
- Replace the second penalty term in AIC and BIC by a term taking into account the number of **models per dimension**.
- Associate to the models $\{V_I : I \in \mathscr{I}\}$ a probability distribution $\pi = \{\pi_I : I \in \mathscr{I}\}$.
- Build the model selection criterion ($K > 1$, see shortly why)

$$\mathrm{MSC}(I) = \left\|y - \widehat{f}_I\right\|_2^2 + K\sigma^2 \left(\sqrt{\dim(V_I)} + \sqrt{2\log(\pi_I^{-1})}\right)^2$$

$$I_{\mathrm{MSC}} \in \operatorname*{Argmin}_{I \in \mathscr{I}} \mathrm{MSC}(I).$$

- Such a choice of penalty seems cryptic but ...
- it ensures that $R(\widehat{f}_{I_{\mathrm{MSC}}})$ is close to $R(\widehat{f}_{I_{\mathrm{oracle}}})$:
  - π is chosen to penalize overly high-dimensional models.

Penalized empirical risk minimization

- Put the same mass on all models with the same sparsity level.
- Choice 1 : $\pi_I = (1 + 1/p)^{-p} p^{-|I|}$, and thus

$$\log(\pi_I^{-1}) \leq 1 + |I| \log(p).$$

- Choice 2 : $\pi_I = (e-1)/(e - e^{-p})(C_p^{|I|})^{-1} e^{-|I|}$, and thus

$$\log(\pi_I^{-1}) \leq \log(e/(e-1)) + 2|I|(2 + \log(p/|I|)).$$

Penalized empirical risk minimization

- Put the same mass on all models with the same sparsity level.
- Choice 1 : $\pi_I = (1+1/p)^{-p} p^{-|I|}$, and thus

$$\log(\pi_I^{-1}) \leq 1 + |I| \log(p).$$

- Choice 2 : $\pi_I = (e-1)/(e-e^{-p})(C_p^{|I|})^{-1} e^{-|I|}$, and thus

$$\log(\pi_I^{-1}) \leq \log(e/(e-1)) + 2|I|(2 + \log(p/|I|)).$$

- Conclusion : $\log(\pi_I^{-1})$ comparable to $\dim(V_I)$ (up to the log factor).
- The $\log(p)$ factor reflects the number of models per sparsity $|I|$: $\log(C_p^{|I|}) \lesssim |I| \log(p/|I|) \leq |I| \log(p)$.
- It can be shown that this factor is unavoidable.
- For $X = \mathrm{Id}$, it can be shown that

$$\mathbb{E}\left[|I_{\mathrm{MSC}}|\right] \asymp \frac{p^{1-K}}{\sqrt{\pi K \log(p)}} \to 0 \text{ if } K > 1.$$

Penalized empirical risk minimization

- Put the same mass on all models with the same sparsity level.
- Choice 1 : $\pi_I = (1 + 1/p)^{-p} p^{-|I|}$, and thus

$$\log(\pi_I^{-1}) \leq 1 + |I| \log(p).$$

- Choice 2 : $\pi_I = (e-1)/(e - e^{-p})(C_p^{|I|})^{-1} e^{-|I|}$, and thus

$$\log(\pi_I^{-1}) \leq \log(e/(e-1)) + 2|I|(2 + \log(p/|I|)).$$

- Conclusion : $\log(\pi_I^{-1})$ comparable to $\dim(V_I)$ (up to the log factor).
- The $\log(p)$ factor reflects the number of models per sparsity $|I|$: $\log(C_p^{|I|}) \lesssim |I| \log(p/|I|) \leq |I| \log(p)$.
- It can be shown that this factor is unavoidable.
- For $X = \mathrm{Id}$, it can be shown that

$$\mathbb{E}\left[|I_{\mathrm{MSC}}|\right] \asymp \frac{p^{1-K}}{\sqrt{\pi K \log(p)}} \to 0 \text{ if } K > 1.$$

Guarantees

Oracle inequality

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon \qquad \mathscr{I} = \bigcup_{\beta \in \mathbb{R}^p} \mathrm{supp}(\beta)$$

$$I_{\mathrm{MSC}} \in \operatorname*{Argmin}_{I \in \mathscr{I}} \left\| y - \widehat{f}_I \right\|_2^2 + K\sigma^2 \left(\sqrt{\dim(V_I)} + \sqrt{2\log(\pi_I^{-1})} \right)^2$$

Oracle inequality

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon \qquad \mathscr{I} = \bigcup_{\beta \in \mathbb{R}^p} \operatorname{supp}(\beta)$$

$$I_{\mathrm{MSC}} \in \operatorname*{Argmin}_{I \in \mathscr{I}} \left\| y - \widehat{f}_I \right\|_2^2 + K\sigma^2 \left(\sqrt{\dim(V_I)} + \sqrt{2\log(\pi_I^{-1})} \right)^2$$

Theorem *There exists a constant* $C > 1$ *depending only on* $K > 1$, *such that*

$$\mathbb{E}\left[\left\| \widehat{f}_{I_{\mathrm{MSC}}} - f^\star \right\|_2^2 \right] \leq C \min_{I \in \mathscr{I}} \Big[\underbrace{\mathbb{E}\left[\left\| \widehat{f}_I - f^\star \right\|_2^2 \right]}_{\text{Risk}} + \underbrace{\sigma^2 \log(\pi_I^{-1})}_{\text{Complexity}} + \sigma^2 \Big].$$

Oracle inequality

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon \qquad \mathscr{I} = \bigcup_{\beta\in\mathbb{R}^p} \mathrm{supp}(\beta)$$

$$I_{\mathrm{MSC}} \in \operatorname*{Argmin}_{I\in\mathscr{I}} \|y - \widehat{f}_I\|_2^2 + K\sigma^2 \left(\sqrt{\dim(V_I)} + \sqrt{2\log(\pi_I^{-1})}\right)^2$$

Theorem *There exists a constant* $C > 1$ *depending only on* $K > 1$, *such that*

$$\mathbb{E}\left[\|\widehat{f}_{I_{\mathrm{MSC}}} - f^\star\|_2^2\right] \le C \min_{I\in\mathscr{I}} \left[\overbrace{\mathbb{E}\left[\|\widehat{f}_I - f^\star\|_2^2\right]}^{\text{Risk}} + \overbrace{\sigma^2\log(\pi_I^{-1})}^{\text{Complexity}} + \sigma^2\right].$$

- Choice 1 : $\pi_I = (1+1/p)^{-p} p^{-|I|}$, and thus

$$\mathbb{E}\left[\|\widehat{f}_{I_{\mathrm{MSC}}} - f^\star\|_2^2\right] \le C \min_{I\in\mathscr{I}} \left[\mathbb{E}\left[\|\widehat{f}_I - f^\star\|_2^2\right] + \overbrace{3\sigma^2|I|\log(p)}^{\text{Complexity}}\right].$$

- Choice 2 : $\pi_I = (e-1)/(e-e^{-p})(C_p^{|I|})^{-1}e^{-|I|}$, and thus

$$\mathbb{E}\left[\|\widehat{f}_{I_{\mathrm{MSC}}} - f^\star\|_2^2\right] \le C \min_{I\in\mathscr{I}} \left[\mathbb{E}\left[\|\widehat{f}_I - f^\star\|_2^2\right] + \overbrace{3\sigma^2|I|\log(p/|I|)}^{\text{Complexity}}\right].$$

- Conclusion : MSC achieves the best tradeoff between the risk and the complexity of the model.

Optimality: minimax risk

$$y = \underbrace{X\beta^\star}_{f^\star \in \mathcal{M}_s} + \varepsilon \qquad \mathcal{M}_s \overset{\text{def}}{=} \bigcup_{I \in \mathscr{I}: |I|=s} \underbrace{\operatorname{Span}(X_I)}_{V_I}, \quad s \le p/2$$

$$\begin{aligned}
\sup_{f^\star \in \mathcal{M}_s} \mathbb{E}\left[\left\| \widehat{f}_{I_{\mathrm{MSC}}} - f^\star \right\|_2^2\right] &\le C \sup_{f^\star \in \mathcal{M}_s} \min_{I \in \mathscr{I}} \left[\mathbb{E}\left[\left\|\widehat{f}_I - f^\star\right\|_2^2\right] + 3\sigma^2 |I| \log(p/|I|)\right] \\
&\le C \sup_{f^\star \in \mathcal{M}_s} \min_{I \in \mathscr{I}: |I|=s} \left[\underbrace{\left\| \operatorname{Proj}_{V_I^\perp} f^\star \right\|_2^2}_{=0} + 4\sigma^2 s \log(p/s)\right] = C'\sigma^2 s \log(p/s).
\end{aligned}$$

Optimality: minimax risk

$$y = \underbrace{X\beta^\star}_{f^\star \in \mathcal{M}_s} + \varepsilon \qquad \mathcal{M}_s \stackrel{\text{def}}{=} \bigcup_{I\in\mathscr{I}:|I|=s} \underbrace{\mathrm{Span}(X_I)}_{V_I}, \quad s \le p/2$$

$$\sup_{f^\star\in\mathcal{M}_s} \mathbb{E}\left[\left\|\widehat{f}_{I_{\mathrm{MSC}}} - f^\star\right\|_2^2\right] \le C \sup_{f^\star\in\mathcal{M}_s} \min_{I\in\mathscr{I}} \left[\mathbb{E}\left[\left\|\widehat{f}_I - f^\star\right\|_2^2\right] + 3\sigma^2|I|\log(p/|I|)\right]$$

$$\le C \sup_{f^\star\in\mathcal{M}_s} \min_{I\in\mathscr{I}:|I|=s} \left[\underbrace{\left\|\mathrm{Proj}_{V_I^\perp} f^\star\right\|_2^2}_{=0} + 4\sigma^2 s\log(p/s)\right] = C'\sigma^2 s\log(p/s).$$

$$\delta_{\min} = \inf_{\beta:\|\beta\|_0=2s} \frac{\|X\beta\|_2}{\|\beta\|_2} \le \sup_{\beta:\|\beta\|_0=2s} \frac{\|X\beta\|_2}{\|\beta\|_2} = \delta_{\max}.$$

Theorem *For any $s \le p/5$, we have*

$$\inf_{\widehat{f}} \sup_{f^\star\in\mathcal{M}_s} \mathbb{E}\left[\left\|\widehat{f} - f^\star\right\|_2^2\right] \ge \frac{e}{8(2e+1)^2}\left(\frac{\delta_{\min}}{\delta_{\max}}\right)^2 \sigma^2 s\log(p/(5s)).$$

Optimality: minimax risk

$$y = \underbrace{X\beta^\star}_{f^\star \in \mathcal{M}_s} + \varepsilon \qquad \mathcal{M}_s \stackrel{\text{def}}{=} \bigcup_{I \in \mathscr{I} : |I| = s} \underbrace{\mathrm{Span}(X_I)}_{V_I}, \quad s \le p/2$$

$$\sup_{f^\star \in \mathcal{M}_s} \mathbb{E}\left[\left\|\widehat{f}_{I_{\mathrm{MSC}}} - f^\star\right\|_2^2\right] \le C \sup_{f^\star \in \mathcal{M}_s} \min_{I \in \mathscr{I}} \left[\mathbb{E}\left[\left\|\widehat{f}_I - f^\star\right\|_2^2\right] + 3\sigma^2 |I| \log(p/|I|)\right]$$

$$\le C \sup_{f^\star \in \mathcal{M}_s} \min_{I \in \mathscr{I} : |I| = s} \left[\underbrace{\left\|\mathrm{Proj}_{V_I^\perp} f^\star\right\|_2^2}_{=0} + 4\sigma^2 s \log(p/s)\right] = C' \sigma^2 s \log(p/s).$$

$$\delta_{\min} = \inf_{\beta : \|\beta\|_0 = 2s} \frac{\|X\beta\|_2}{\|\beta\|_2} \le \sup_{\beta : \|\beta\|_0 = 2s} \frac{\|X\beta\|_2}{\|\beta\|_2} = \delta_{\max}.$$

Theorem *For any $s \le p/5$, we have*

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{M}_s} \mathbb{E}\left[\left\|\widehat{f} - f^\star\right\|_2^2\right] \ge \frac{e}{8(2e+1)^2} \left(\frac{\delta_{\min}}{\delta_{\max}}\right)^2 \sigma^2 s \log(p/(5s)).$$

Upper and lower bound are of the same order $\sim s \log(n/s) \Rightarrow \widehat{f}_{I_{\mathrm{MSC}}}$ is optimal.

Optimality: minimax risk

$$y = \underbrace{X\beta^\star}_{f^\star \in \mathcal{M}_s} + \varepsilon \qquad \mathcal{M}_s \stackrel{\text{def}}{=} \bigcup_{I \in \mathscr{I} : |I| = s} \underbrace{\mathrm{Span}(X_I)}_{V_I}, \quad s \le p/2$$

$$\sup_{f^\star \in \mathcal{M}_s} \mathbb{E}\left[\|\widehat{f}_{I_{\mathrm{MSC}}} - f^\star\|_2^2\right] \le C \sup_{f^\star \in \mathcal{M}_s} \min_{I \in \mathscr{I}} \left[\mathbb{E}\left[\|\widehat{f}_I - f^\star\|_2^2\right] + 3\sigma^2 |I| \log(p/|I|)\right]$$

$$\le C \sup_{f^\star \in \mathcal{M}_s} \min_{I \in \mathscr{I} : |I| = s} \left[\underbrace{\|\mathrm{Proj}_{V_I^\perp} f^\star\|_2^2}_{=0} + 4\sigma^2 s \log(p/s)\right] = C'\sigma^2 s \log(p/s).$$

$$\delta_{\min} = \inf_{\beta : \|\beta\|_0 = 2s} \frac{\|X\beta\|_2}{\|\beta\|_2} \le \sup_{\beta : \|\beta\|_0 = 2s} \frac{\|X\beta\|_2}{\|\beta\|_2} = \delta_{\max}.$$

Theorem *For any $s \le p/5$, we have*

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{M}_s} \mathbb{E}\left[\|\widehat{f} - f^\star\|_2^2\right] \ge \frac{e}{8(2e+1)^2} \left(\frac{\delta_{\min}}{\delta_{\max}}\right)^2 \sigma^2 s \log(p/(5s)).$$

Upper and lower bound are of the same order $\sim s\log(n/s) \Rightarrow \widehat{f}_{I_{\mathrm{MSC}}}$ is optimal.

Rmainder captures a phase transition phenomenon between good and poor estimation, i.e. $n \gtrsim s\log(p)$.

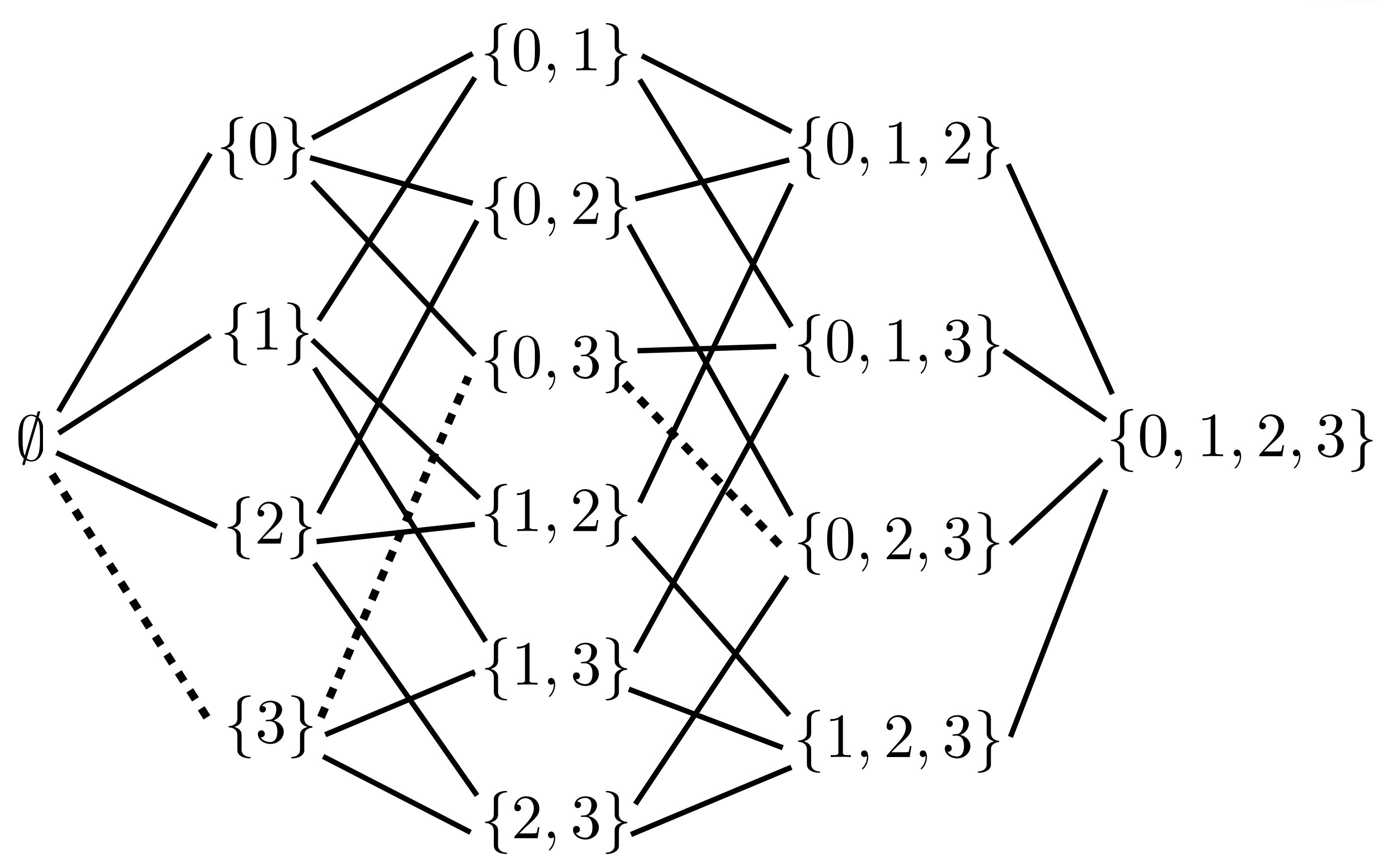
Computational issues

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon \qquad \mathscr{I} = \bigcup_{\beta\in\mathbb{R}^p} \mathrm{supp}(\beta)$$

$$|\mathscr{I}| = 2^p$$

$$I_{\mathrm{MSC}} \in \operatorname*{Argmin}_{I\in\mathscr{I}} \|y - \widehat{f}_I\|_2^2 + K\sigma^2 \left(\sqrt{\dim(V_I)} + \sqrt{2\log(\pi_I^{-1})}\right)^2$$

This cannot be implemented in practice except for the orthogonal case (requires exhaustive search) : NP-hard problem.



Computational issues

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon \qquad \mathscr{I} = \bigcup_{\beta \in \mathbb{R}^p} \mathrm{supp}(\beta)$$

$$|\mathscr{I}| = 2^p$$

$$I_{\mathrm{MSC}} \in \operatorname*{Argmin}_{I \in \mathscr{I}} \|y - \widehat{f}_I\|_2^2 + K\sigma^2 \left(\sqrt{\dim(V_I)} + \sqrt{2\log(\pi_I^{-1})}\right)^2$$

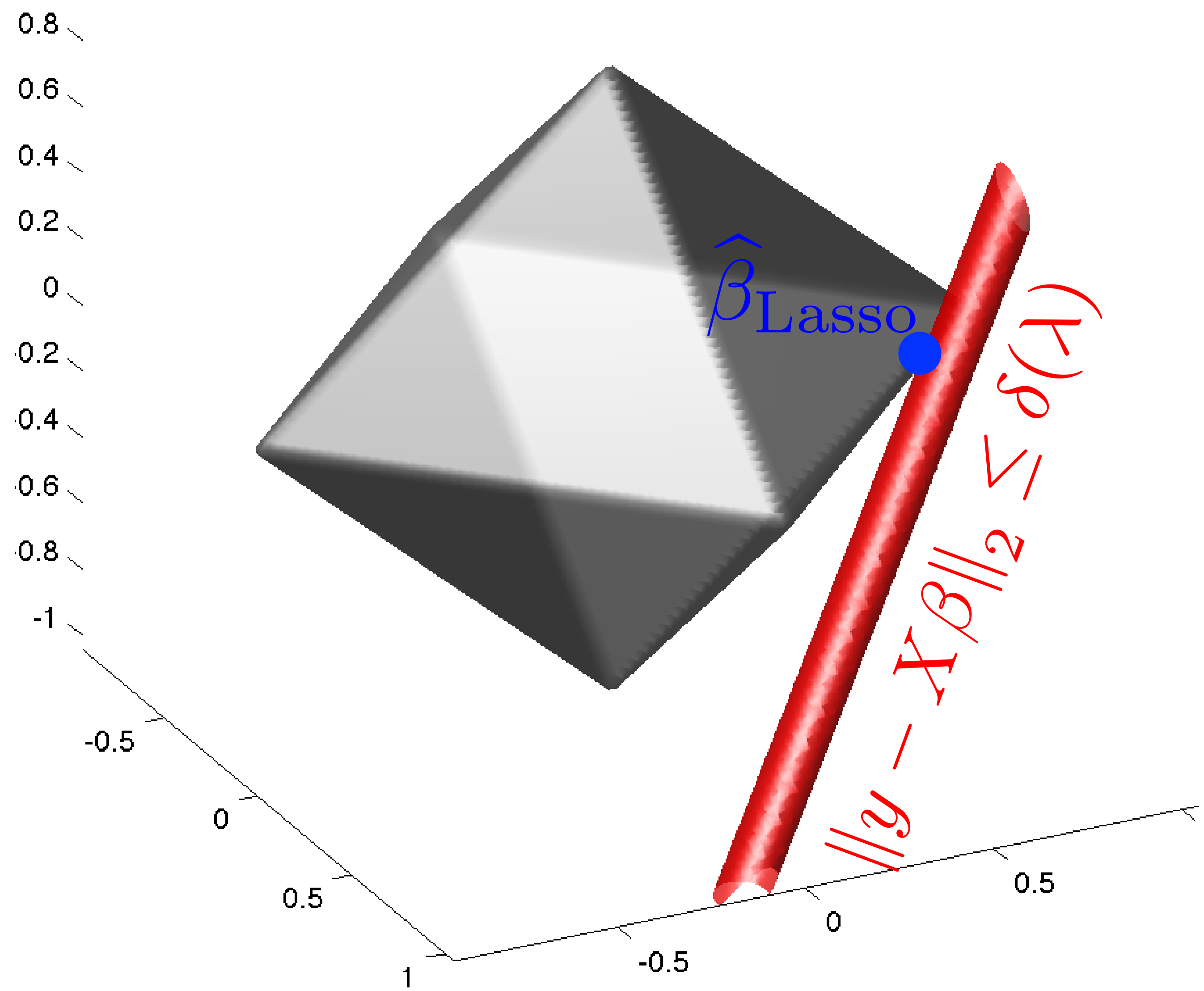
This cannot be implemented in practice except for the orthogonal case (requires exhaustive search) : NP-hard problem.

Can we have a polynomial-time procedure to do so with the same performance guarantees ?

Convex relaxation: Lasso

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon$$

$$\boxed{\widehat{f}_{\text{Lasso}} = X \operatorname*{Argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1}$$



Lasso oracle inequality

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon$$

$$\boxed{\widehat{f}_{\text{Lasso}} = X \operatorname*{Argmin}_{\beta\in\mathbb{R}^p} \left\|y - X\beta\right\|_2^2 + \lambda\left\|\beta\right\|_1}$$

$$\Upsilon(I,c) \stackrel{\text{def}}{=} \inf_{\left\{\alpha\in\mathbb{R}^p \,:\, \|\alpha_{I^c}\|_1 < c\|\alpha_I\|_1\right\}} \frac{\sqrt{|I|}\|X\alpha\|_2}{\|\alpha_I\|_1}$$

A measure of restricted ill-conditioning

Lasso oracle inequality

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon \qquad \boxed{\widehat{f}_{\text{Lasso}} = X \operatorname*{Argmin}_{\beta \in \mathbb{R}^p} \left\| y - X\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1}$$

$$\Upsilon\left(I, c\right) \overset{\text{def}}{=} \inf_{\left\{\alpha \in \mathbb{R}^p \,:\, \|\alpha_{I^c}\|_1 < c\|\alpha_I\|_1\right\}} \frac{\sqrt{|I|}\|X\alpha\|_2}{\|\alpha_I\|_1}$$

A measure of restricted ill-conditioning

Theorem *Assume that the columns of X have unit-norm. Then, $\forall \delta > 1$, the Lasso with*

$$\lambda = 2\sigma\sqrt{2\delta\log(2p)},$$

obeys with probability at least $1 - 2(2p)^{1-\delta}$

Complexity

$$\left\|\widehat{f}_{\text{Lasso}} - f^\star\right\|_2^2 \leq \min_{\substack{I \in \mathscr{I} \\ \operatorname{supp}(\beta) = I}} \left[\left\|X\beta - f^\star\right\|_2^2 + \boxed{\frac{9\sigma^2\delta^2}{\Upsilon(I,3)^2}|I|\log(2p)}\right].$$

Lasso oracle inequality

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon$$

$$\boxed{\widehat{f}_{\text{Lasso}} = X \operatorname*{Argmin}_{\beta \in \mathbb{R}^p} \left\| y - X\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1}$$

$$\Upsilon(I, c) \overset{\text{def}}{=} \inf_{\left\{\alpha \in \mathbb{R}^p \,:\, \|\alpha_{I^c}\|_1 < c \|\alpha_I\|_1\right\}} \frac{\sqrt{|I|}\|X\alpha\|_2}{\|\alpha_I\|_1}$$

A measure of restricted ill-conditioning

Theorem *Assume that the columns of X have unit-norm. Then, $\forall \delta > 1$, the Lasso with*

$$\lambda = 2\sigma\sqrt{2\delta\log(2p)},$$

obeys with probability at least $1 - 2(2p)^{1-\delta}$

Complexity

$$\left\| \widehat{f}_{\text{Lasso}} - f^\star \right\|_2^2 \leq \min_{\substack{I \in \mathscr{I} \\ \operatorname{supp}(\beta) = I}} \left[\left\| X\beta - f^\star \right\|_2^2 + \frac{9\sigma^2\delta^2}{\Upsilon(I,3)^2} |I| \log(2p) \right].$$

$$\mathbb{E}\left[\left\| \widehat{f}_{I_{\text{MSC}}} - f^\star \right\|_2^2 \right] \leq C \min_{I \in \mathscr{I}} \left[\mathbb{E}\left[\left\| \widehat{f}_I - f^\star \right\|_2^2 \right] + 3\sigma^2 |I| \log(p) \right].$$

Remainder term of the same order as MSC while Lasso is implementable.

The OI is sharp for Lasso ($C = 1$).

Implementation of Lasso

$$y = \underbrace{X\beta^{\star}}_{f^{\star}} + \varepsilon \qquad \widehat{f}_{\mathrm{Lasso}} = X \operatorname*{Argmin}_{\beta \in \mathbb{R}^p} \left\| y - X\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1$$

See MC on non-smooth optimization.

Aggregation by exponential weighting

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon$$

- Idea : rather than selecting one model, take the best of each model by averaging.
- The Exponential Weighted Aggregation (EWA) estimator is

$$F(\beta) \stackrel{\text{def}}{=} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \qquad \widehat{\beta}_{\text{EWA}} = \int_{\mathbb{R}^p} \beta\mu(\beta)d\beta$$

$$\mu(\beta) = \frac{\exp\left(-F(\beta)/\tau\right)}{\int_{\mathbb{R}^p} \exp\left(-F(\alpha)/\tau\right)d\alpha},$$

Aggregation by exponential weighting

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon$$

- Idea : rather than selecting one model, take the best of each model by averaging.
- The Exponential Weighted Aggregation (EWA) estimator is

$$F(\beta) \stackrel{\text{def}}{=} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \qquad \widehat{\beta}_{\text{EWA}} = \int_{\mathbb{R}^p} \beta\mu(\beta)d\beta$$

$$\mu(\beta) = \frac{\exp\left(-F(\beta)/\tau\right)}{\int_{\mathbb{R}^p} \exp\left(-F(\alpha)/\tau\right)d\alpha},$$

Involves solving an **integration** problem.

EWA oracle inequality

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon$$

$$\boxed{\widehat{\beta}_{\mathrm{EWA}} = \int_{\mathbb{R}^p} \beta \mu(\beta) d\beta}$$

$$\Upsilon(I, c) \stackrel{\text{def}}{=} \inf_{\left\{\alpha \in \mathbb{R}^p \,:\, \|\alpha_{I^c}\|_1 < c \|\alpha_I\|_1\right\}} \frac{\sqrt{|I|}\|X\alpha\|_2}{\|\alpha_I\|_1}$$

A measure of restricted ill-conditioning

EWA oracle inequality

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon$$

$$\boxed{\widehat{\beta}_{\mathrm{EWA}} = \int_{\mathbb{R}^p} \beta \mu(\beta) d\beta}$$

$$\Upsilon(I,c) \stackrel{\text{def}}{=} \inf_{\left\{\alpha \in \mathbb{R}^p \,:\, \|\alpha_{I^c}\|_1 < c\|\alpha_I\|_1\right\}} \frac{\sqrt{|I|}\|X\alpha\|_2}{\|\alpha_I\|_1}$$ A measure of restricted ill-conditioning

Theorem *Assume that the columns of X have unit-norm. Then, $\forall \delta > 1$, and for some absolute constant $C' > 0$, the EWA with*

$$\lambda = 2\sigma\sqrt{2\delta \log(2p)} \quad \textit{and} \quad \tau = O(1/p)$$

obeys with probability at least $1 - 2(2p)^{1-\delta}$

$$\left\|\widehat{f}_{\mathrm{Lasso}} - f^\star\right\|_2^2 \leq \min_{\substack{I \in \mathscr{I} \\ \mathrm{supp}(\beta) = I}} \left[\left\|X\beta - f^\star\right\|_2^2 + \underbrace{\frac{C'\sigma^2\delta^2}{\Upsilon(I,3)^2}|I| \log(2p)}_{\text{Complexity}}\right],$$

EWA oracle inequality

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon$$

$$\widehat{\beta}_{\mathrm{EWA}} = \int_{\mathbb{R}^p} \beta\mu(\beta)d\beta$$

$$\Upsilon(I,c) \stackrel{\mathrm{def}}{=} \inf_{\left\{\alpha\in\mathbb{R}^p \,:\, \|\alpha_{I^c}\|_1 < c\|\alpha_I\|_1\right\}} \frac{\sqrt{|I|}\|X\alpha\|_2}{\|\alpha_I\|_1}$$

A measure of restricted ill-conditioning

Theorem *Assume that the columns of* X *have unit-norm. Then,* $\forall\delta > 1$*, and for some absolute constant* $C' > 0$*, the EWA with*

$$\lambda = 2\sigma\sqrt{2\delta\log(2p)} \quad \textit{and} \quad \tau = O(1/p)$$

obeys with probability at least $1 - 2(2p)^{1-\delta}$

Complexity

$$\left\|\widehat{f}_{\mathrm{Lasso}} - f^\star\right\|_2^2 \leq \min_{\substack{I\in\mathscr{I} \\ \mathrm{supp}(\beta)=I}} \left[\left\|X\beta - f^\star\right\|_2^2 + \boxed{\frac{C'\sigma^2\delta^2}{\Upsilon(I,3)^2}|I|\log(2p)}\right],$$

$$\mathbb{E}\left[\left\|\widehat{f}_{I_{\mathrm{MSC}}} - f^\star\right\|_2^2\right] \leq C\min_{I\in\mathscr{I}}\left[\mathbb{E}\left[\left\|\widehat{f}_I - f^\star\right\|_2^2\right] + 3\sigma^2|I|\log(p)\right].$$

Remainder term of the same order as MSC and Lasso.

The OI is sharp for EWA as well.

Implementation of EWA

$$y = \underbrace{X\beta^\star}_{f^\star} + \varepsilon$$

$$\boxed{\widehat{\beta}_{\text{EWA}} = \int_{\mathbb{R}^p} \beta \mu(\beta) d\beta}$$

- An integration problem in high dimension.
- Very challenging.
- MC sampling through e.g. SDE's.
- Maybe another MC.

Take-away messages

- Model selection benefits from blessings of dimensionality.
- Several model selection criteria in the literature.
- AIC and BIC are not good when the number of models is very large.
- MSC is much better and enjoys nice guarantees, but is NP-hard.
- Lasso is implementable (convex program) while enjoying the same guarantees.
- So does EWA but necessitates even more sophisticated sampling algorithms.
- We gave principles for estimation but other tasks as well: classification, machine learning.

https://fadili.users.greyc.fr/

Thanks
Any questions ?