gipsa-lab

Grenoble | images | parole | signal | automatique | laboratoire

A Concentration of Measure & Random Matrix Perspective to Machine Learning

Random Matrix Seminar – Mathematical Institute, Oxford.

Cosme Louart, Romain Couillet

GIPSA-lab, INP Grenoble; List-CEA

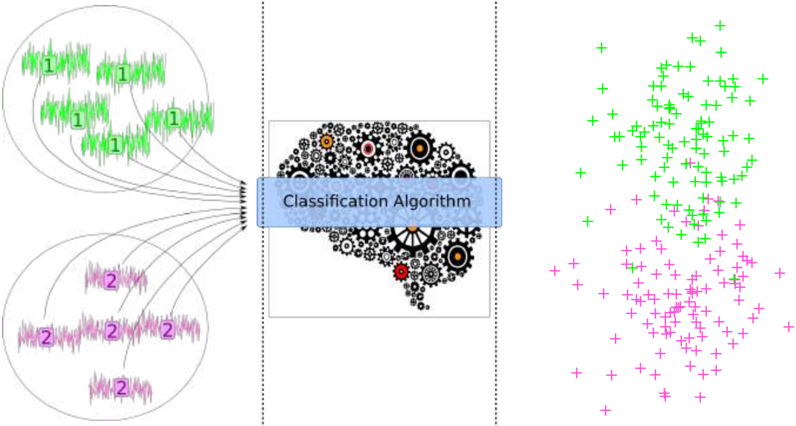
18/02/2020

cnrs Grenoble INP UNIVERSITÉ Grenoble Alpes UMR 5216

Position of the problem

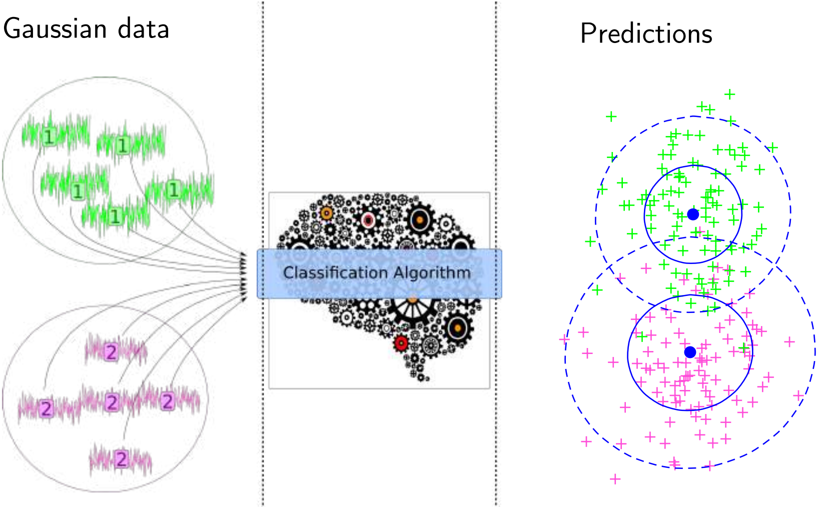
Gaussian data

Classification output



How to predict performances in high dimension ?

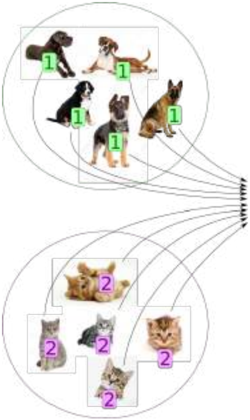
Position of the problem

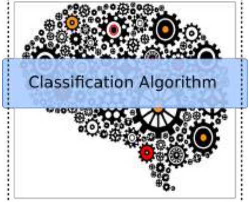


$\Longrightarrow$ Resort to Random Matrix Theory tools

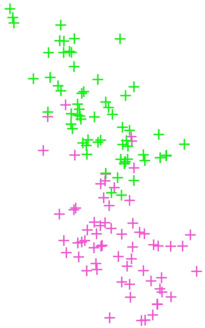
Position of the problem

Real data





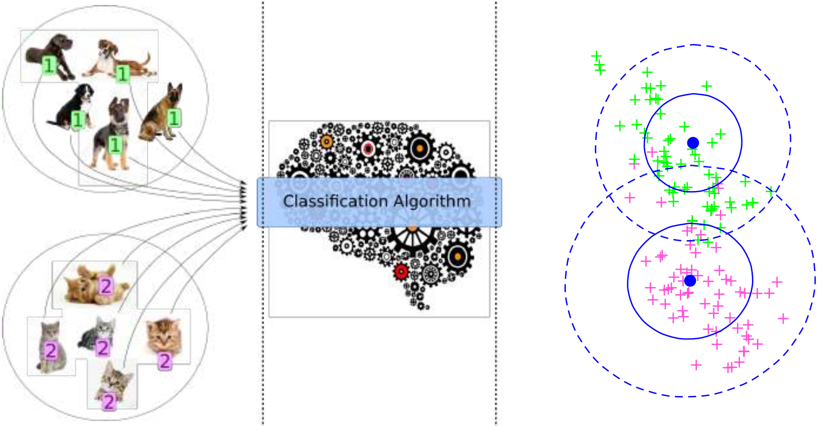
Classification output



How to predict performances in realistic settings ?

Position of the problem

Real data | Gaussian Predictions



$\Longrightarrow$ Resort to RMT + Concentration of measure hypotheses

Table of Contents

Table of contents

Classical study of singular values of rectangular RM

$X = (x_1, \ldots, x_n) \in \mathcal{M}_{p,n}$, spectral distribution of $\frac{1}{n}XX^T$:

$$\frac{1}{p} \sum_{\lambda \in \mathrm{Sp}(\frac{1}{n}XX^T)} \delta_\lambda$$

Classical Hypothesis

- X has i.i.d entries with bounded variance
- $X = C^{\frac{1}{2}}Z$, $Z \sim \mathcal{N}(0, I_n)$.

Classical conclusions

- Weak convergence of the spectral distribution to the Marcenko-Pastur law

Question : Can we find relaxed hypotheses and control the speed of convergence ?

With the concentration of measure theory (CMT)

Hypothesis of CMT

1. For all 1-Lipschitz maps $f : \mathcal{M}_{p,n} \to \mathbb{R}$:

$$\forall t > 0 : \quad \mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) \leq 2e^{-t^2/2}$$

   (Independently on p and n !)

2. The column of X are i.i.d.

Remarks

- **(Cons)** Implies all the moments are bounded
- **(Pros)** True if the columns are Lipschitz transformations of a Gaussian vector $Z \sim \mathcal{N}(0, I_p)$.
  $\longrightarrow$ dependence between entries of a column possibly complex

With the concentration of measure theory (CMT)

Conclusions on the spectral distribution

- Noting $Q(z) = (\frac{1}{n}XX^T + zI_p)^{-1}$, the resolvent of $\frac{1}{n}XX^T$, $(\frac{1}{p}\operatorname{Tr}(Q(z))$: *Stieltjes transform*)

$$\exists C, c \underset{p,n\to\infty}{=} O(1) :$$
$$\forall t > 0 : \quad \mathbb{P}\left(\left| \operatorname{Tr}(AQ(z)) - \operatorname{Tr}(A\tilde{Q}) \right| \geq t \right) \leq C \exp\left(-\frac{nt^2}{c\|A\|_1^2} \right)$$

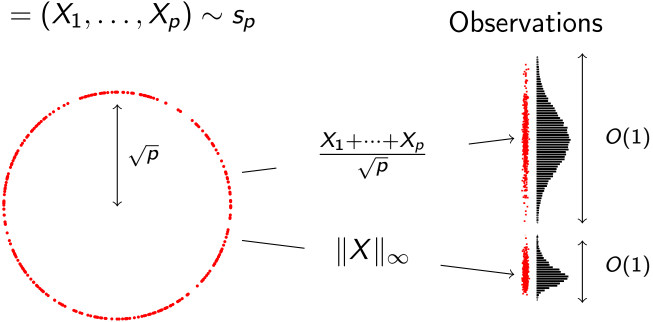
where $\tilde{Q} \in \mathcal{M}_p$ is a deterministic equivalent of Q

(if $A = \frac{1}{p} I_n$: convergence of the Stieltjes transform of spectral dist. of $\frac{1}{nXX^T}$)

Table of contents

Concentration of Measure Phenomenon[1]

$$X = (X_1, \ldots, X_p) \sim s_p$$



Distribution diameter $= \mathbb{E}[\|Z - \mathbb{E}Z\|] \underset{p\to\infty}{=} O(\sqrt{p})$

Observable diameter $\underset{p\to\infty}{=} O(1)$

[1] Ledoux - 2001 : The concentration of measure phenomenon

Setting

$(E, \|\cdot\|)$, a normed vector space, $Z_{n,p} \in E$, a random vector

- $(\mathbb{R}^p, \|\cdot\|)$, with $\|x\| = \sqrt{\sum_{i=1}^p x_i^2}$
- $(\mathcal{M}_{p,n}, \|\cdot\|_F)$ with $\|M\|_F = \sqrt{\mathrm{Tr}(MM^T)} = \sqrt{\sum_{\substack{1\leq i\leq p \\ 1\leq j\leq n}} M_{i,j}^2}$
- $(\mathcal{M}_{p,n}, \|\cdot\|)$ with $\|M\| = \sup_{\|x\|\leq 1} \|Mx\|$

Definition of concentration

if $\exists C, c > 0,\ (\sigma_{p,n})_{p,n\in\mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}^2} \mid \forall n, p \in \mathbb{N}$,
$\forall f : (E, \|\cdot\|) \to (\mathbb{R}, |\cdot|)$ 1-Lipschitz:

$$\forall t > 0\ :\ \mathbb{P}\left(|f(Z_{n,p}) - \mathbb{E}[f(Z_{n,p})]| \geq t\right) \leq Ce^{-(t/c\sigma_{p,n})^2},$$

we note $Z \propto \mathcal{E}_2(\sigma)$

Fundamental example of the Theory:

$Z \in \mathbb{R}^p$, if $Z \sim \text{Unif}(\sqrt{p}\mathcal{S}^{p-1})$, $Z \sim \text{Unif}(\mathcal{B}_{\mathbb{R}^p}(0, \sqrt{p}))$ or $Z \sim \mathcal{N}(0, I_p)$:
$\forall f : E \to \mathbb{R}$ 1-Lipschitz :

$$\forall t > 0 \ : \ \mathbb{P}\left(\left|f(Z) - \mathbb{E}[f(Z')]\right| \geq t\right) \leq 2e^{-t^2/2},$$

Choosing $C = 2, c = \sqrt{2}, \sigma_p = 1$:

$$Z \propto \mathcal{E}_2(1) \text{ (Independent of p !).}$$

$\longrightarrow$Standard Hypothesis : $Z \propto \mathcal{E}_2$

Notion of deterministic equivalent

- if $C, c > 0$, $(\sigma_{p,n})_{p,n\in\mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}^2} \mid \forall n, p \in \mathbb{N}$,
  $\forall f : (E, \|\cdot\|) \to (\mathbb{R}, |\cdot|)$ 1-Lipschitz :

$$\boxed{\forall t > 0 \ : \ \mathbb{P}\left(|f(Z) - \mathbb{E}[f(Z)]| \geq t\right) \leq Ce^{-(t/c\sigma_{p,n})^2},}$$

  Notation: $Z \propto \mathcal{E}_2(\sigma)$

- In particular, if $\exists \tilde{Z} \in E \mid \forall u : E \to \mathbb{R}$ 1-Lipschitz and linear :

$$\boxed{\forall t > 0 \ : \ \mathbb{P}\left(\left|u(Z - \tilde{Z})\right| \geq t\right) \leq Ce^{-(t/c\sigma_{p,n})^2},}$$

  Notation: $Z \in \tilde{Z} \pm \mathcal{E}_2(\sigma)$
  $\tilde{Z}$: Deterministic equivalent of Z.
  Of course: $Z \propto \mathcal{E}_2(\sigma) \Longrightarrow Z \in \mathbb{E}[Z] \pm \mathcal{E}_2(\sigma)$

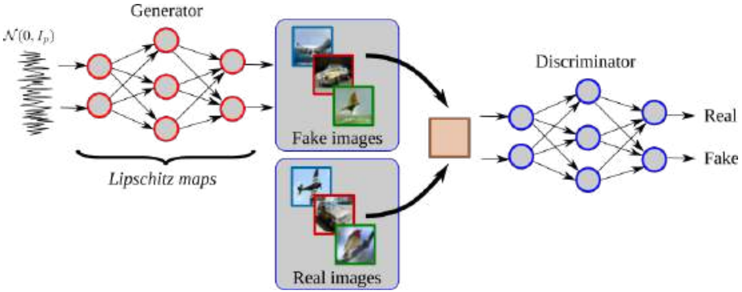
Strategy for the study of $Q = (\frac{1}{n}XX^T + \gamma I_p)^{-1}$

1. $Z \propto \mathcal{E}_2$ in $(\mathbb{R}^p, \|\cdot\|)$
   $\implies Q \propto \mathcal{E}_2(1/\sqrt{n})$ in $(\mathcal{M}_{p,n}, \|\cdot\|_F)$
   $\implies Q \in \mathbb{E}[Q] \pm \mathcal{E}_2(1/\sqrt{n})$ in $(\mathcal{M}_{p,n}, \|\cdot\|_F)$
2. $\exists \tilde{Q} \in \mathcal{M}_p, \ \|\mathbb{E}[Q] - \tilde{Q}\| = O(1/\sqrt{n})$
   $\implies Q \in \tilde{Q} \pm \mathcal{E}_2(1/\sqrt{n})$ in $(\mathcal{M}_{p,n}, \|\cdot\|)$

How to build new concentrated random vectors ?

- If $Z \propto \mathcal{E}_2(\sigma)$ and $f : E \to E$ λ-Lipschitz, $f(Z) \propto \mathcal{E}_2(\lambda\sigma)$
- No simple way to set the concentration of $(Z_1, \ldots, Z_p)$ if $Z_1, \ldots, Z_p \propto \mathcal{E}_2(\sigma)$. Two possibilities:
  1. Z_1, Z_2, independent then: $(Z_1, Z_2) \propto \mathcal{E}_2(\sigma)$
  2. $(Z_1, Z_2) = f(Z)$ where $Z \propto \mathcal{E}_2(\sigma)$, and f 1-Lipschitz. Then: $(Z_1, Z_2) \propto \mathcal{E}_2(\sigma)$

Realistic images built with GANs are concentrated



$$\textcolor{red}{\text{FAKE IMAGE} = f(Z)}, \quad \text{with } f \text{ } 1-\text{Lipschitz and } Z \sim \mathcal{N}(0, I_p)$$

Characterization with the moments

$$Z \propto \mathcal{E}_2(\sigma) \quad \Longleftrightarrow \quad \left\{ \begin{array}{c} \forall r \geq q, \forall f : E \to \mathbb{R}, \text{1-Lipschitz}, \exists c > 0 : \\ \mathbb{E}\left[|f(Z) - \mathbb{E}[f(Z)]|^r\right] \leq C \left(\frac{r}{q}\right)^{\frac{r}{q}} (c\sigma)^r \end{array} \right.$$

Proof ($f \equiv f(Z)$ and $\bar{f} = \mathbb{E}[f(Z)]$):

$\Rightarrow$ Fubini:

$$\begin{aligned} \mathbb{E}\left[\left|f - \bar{f}\right|^r\right] &= \int_Z \left(\int_0^\infty \mathbb{1}_{t \leq \left|f - \bar{f}\right|^r} dt\right) dZ \\ &= \int_0^\infty \mathbb{P}\left(\left|f - \bar{f}\right|^r \geq t\right) dt \\ &\leq \int_0^\infty C e^{-t^{\frac{q}{r}}/(c\sigma)^q} dt \ldots \leq C' \left(\frac{r}{q}\right)^{\frac{r}{q}} \sigma^r \end{aligned}$$

$\Leftarrow$ Markov inequality:

$$\mathbb{P}\left(\left|f - \bar{f}\right| \geq t\right) \leq \frac{\mathbb{E}\left[\left|f - \bar{f}\right|^r\right]}{t^r} \leq C \left(\frac{r}{q}\right)^{\frac{r}{q}} \left(\frac{c\sigma}{t}\right)^r,$$

with $r = \frac{qt^q}{e(c\sigma)^q} \geq q : \mathbb{P}\left(\left|f - \bar{f}\right| \geq t\right) \leq Ce^{-(t/c\sigma)^q/e}$.

Table of contents

Key result : Control of the norm

- Infinite norm :

$$\mathbb{P}\left(\|Z-\tilde{Z}\|_\infty \geq t\right) = \mathbb{P}\left(\sup_{1\leq i\leq p} e_i^T(Z-\tilde{Z}) \geq t\right)$$
$$\leq p \sup_{1\leq i\leq p} \mathbb{P}\left(e_i^T(Z-\tilde{Z}) \geq t\right)$$
$$\leq Ce^{\log p-(t/c\sigma)^2} \leq C'e^{-(t/\sigma\sqrt{\log(p)})^2},$$

- For the general case, use of "ε-nets". If $\exists H \subset (E^*, \|\cdot\|_*)$ $|$

$$\forall z \in E: \ \|z\| = \sup_{f\in H} f(z).$$

$$Z \in \tilde{Z} \pm C\mathcal{E}_q(\sigma) \implies \left\|Z-\tilde{Z}\right\| \in 0 \pm 8^{\dim(\mathrm{Vect}(H))} C\mathcal{E}_q(2\sigma)$$

on $(\mathbb{R}^p, \|\cdot\|)$, $H=\mathbb{R}^p$, and $\dim(\mathrm{Vect}(H)) = p$

Norm degree

Degree of a subset $H \subset E^*$ and of a norm

- $\eta_H = \log(\#H)$ if H is finite
- $\eta_H = \dim(\mathrm{Vect}(H))$ if H is infinite

Degree of a norm

- $\eta_{\|\cdot\|} = \inf \left\{ \eta_H, H \subset E^* \mid \forall x \in E, \|x\| = \sup_{f \in H} f(x) \right\}$

Example

- $\eta\left(\mathbb{R}^p, \|\cdot\|_\infty\right) = \log(p)$
- $\eta\left(\mathbb{R}^p, \|\cdot\|\right) = p$
- $\eta\left(\mathcal{M}_{p,n}, \|\cdot\|\right) = n + p$
- $\eta\left(\mathcal{M}_{p,n}, \|\cdot\|_F\right) = np.$

Concentration of the norm

If $Z \in \tilde{Z} \pm C\mathcal{E}_q(\sigma)$:

$$\left\| Z - \tilde{Z} \right\| \in 0 \pm \mathcal{E}_q(\sigma \sqrt{\eta_{\|\cdot\|}}) \quad \text{and} \quad \mathbb{E} \left\| Z - \tilde{Z} \right\| = O\left(\sigma \sqrt{\eta_{\|\cdot\|}}\right).$$

Example $Z \in \mathbb{R}^p$, $X \in \mathcal{M}_{p,n}$

- if $Z \in \tilde{Z} \pm \mathcal{E}_2 : \mathbb{E} \|Z\| \leq \|\tilde{Z}\| + C\sqrt{p}$
- if $X \in \tilde{X} \pm \mathcal{E}_2 : \mathbb{E} \|X\| \leq \|\tilde{X}\| + C\sqrt{p+n}$,
- if $X \in \tilde{X} \pm \mathcal{E}_2 : \mathbb{E} \|X\|_F \leq \|\tilde{X}\|_F + C\sqrt{pn}$.

Table of contents

Concentration of the sum and the product

If $(X, Y) \in \mathcal{E}_2(\sigma)$ (X,Y independent or $(X,Y) = f(Z)$, $Z \in \mathcal{E}_2(\sigma)$):

- $X + Y \in \mathcal{E}_2(\sigma)$
- $(X - \tilde{X})(Y - \tilde{Y})$

$$\propto \mathcal{E}_1\left(\sigma^2\right) + \mathcal{E}_2\left(\sigma^2\sqrt{\eta_{\|\cdot\|'}}\right) \quad \text{in} \quad (E, \|\cdot\|)$$

where $\forall x, y \in \mathcal{E}$ $\|xy\| \leq \|x\|'\|y\|$ (usually $\|x\|' \leq \|x\|$).

Example $X \in \mathcal{M}_{p,n}$, $Y, Z \in \mathbb{R}^p$, $Y, Z, X \propto \mathcal{E}_2$

- $\frac{XX^T}{n} \propto \mathcal{E}_2\left(\frac{\sqrt{n+p}}{n}\right) + \mathcal{E}_1\left(\frac{1}{n}\right)$ in $(\mathcal{M}_{p,n}, \|\cdot\|_F)$
- $Y \odot Z \propto \mathcal{E}_2(\sqrt{\log p}) + \mathcal{E}_1$ in $(\mathbb{R}^p, \|\cdot\|)$

Hanson Wright-like results

Classical Theorem[2]

If $Z_1, \dots, Z_p \in \mathbb{R}$, independent, $\forall i, Z_i \propto \mathcal{E}_2$, $\mathbb{E}[Z_i] = 0$:

$$\pi \equiv \mathbb{P}\left(\left|Z^T AZ - \mathbb{E}Z^T AZ\right| \geq t\right) \leq C \exp\left(-c \min\left(\left(\frac{t}{\|A\|_F}\right)^2, \frac{t}{\|A\|}\right)\right)$$

With the Concentration of the measure phenomenon

If $Z = (Z_1, \dots, Z_p) \propto \mathcal{E}_2$, $\mathbb{E}[Z_i] = 0$ two results:

1. $\pi \leq C \exp\left(-c \min\left(\left(\frac{t}{\sqrt{p}\|A\|}\right)^2, \frac{t}{\|A\|}\right)\right)$
2. $\pi \leq C \exp\left(-c \min\left(\left(\frac{t}{\sqrt{\log p}\|A\|}\right)^2, \frac{t}{\|A\|_F}\right)\right)$

[2] Roman Vershynin - High-Dimensional Probability

Proofs

$A = S_+ - S_- + R$ with $S_+, S_- \geq 0$ and R antisymmetric
$\Rightarrow$ enough to prove the result for $A \geq 0$

1. $\boxed{Z^T AZ \propto \mathcal{E}_2\left(\sqrt{p}\,\|A\|\right) + \mathcal{E}_1\left(\|A\|\right)}$:
   $Z^T AZ = \|A^{1/2}Z\|^2$ where $A^{1/2}Z \propto \mathcal{E}_2(\|A\|^{1/2})$ and $\mathbb{E}[\|A^{1/2}Z\|] = \sqrt{n}\|A\|^{1/2}$.

2. $\boxed{Z^T AZ \propto \mathcal{E}_2\left(\sqrt{\log p}\|A\|_F\right) + \mathcal{E}_1\left(\|A\|_F\right)}$
   $A = P^{-1}\Lambda P$, with $\Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_p)$
   $Y \equiv PZ \propto \mathcal{E}_2$ (since $\|P\| = 1$) and:

$$Z^T AZ = (Y \odot Y)^T \lambda$$

   But $Y \odot Y \propto \mathcal{E}_2(\sqrt{\log p}) \pm \mathcal{E}_1$
   $\rightarrow$ we conclude since $\|\lambda\| = \|A\|_F$

Table of contents

Position of the problem

Data matrix $X = (x_1, \dots, x_n) \in \mathcal{M}_{p,n}$,

Hypothesis:

- $p = O(n)$ and $n = O(p)$
- $X \propto \mathcal{E}_2$
- $\|\mathbb{E}[X]\| = O(\sqrt{n})$

Strategy:

1. concentration of the resolvent: $Q = Q(z) = \left(\frac{1}{n}XX^T + zI_p\right)^{-1}$,
2. computable *deterministic equivalent* $\tilde{Q}$,
3. spectral dist. of $\frac{1}{n}XX^T$ from estimation of Stieltjes transf:

$$m(z) = \frac{1}{p}\operatorname{Tr}(Q(z)).$$

1 - Concentration of $Q = Q(z) = \left(\frac{1}{n}XX^T + zI_p\right)^{-1}$

$Q = f(X)$ with f $O(\frac{1}{\sqrt{n}})$-Lipschitz and $X \propto \mathcal{E}_2$thus:

$$Q \in \mathbb{E}[Q] \pm \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right)$$

2 - Choice of a computable deterministic equivalent

- naive choice : $\tilde{Q} = (\Sigma + zI_p)^{-1} \to$ wrong!
- clever choice : $\tilde{Q} = (\frac{\Sigma}{1+\delta} + zI_p)^{-1}$with
  1. $\delta = \frac{1}{n}\operatorname{Tr}(\Sigma\mathbb{E}[Q]) \to$ not directly computable !
  2. δ solution to:

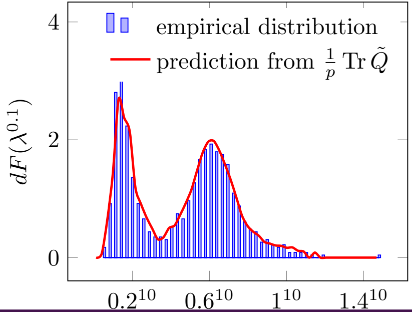
$$\delta = \frac{1}{n}\operatorname{Tr}\left(\Sigma\left(\frac{\Sigma}{1+\delta} + zI_p\right)^{-1}\right)$$

Conclusion for the spectral distribution of $\frac{1}{n}XX^T$

Theorem

$Q \in \tilde{Q} \pm \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right)$ in $(\mathcal{M}_{p,n}, \|\cdot\|)$in particular, $\forall z > 1$:

$$\mathbb{P}\left(\left|m(z) - \frac{1}{p}\mathrm{Tr}(\tilde{Q}(z))\right| \geq t\right) \leq Ce^{-(nt/c)^2}, \quad \text{for} \quad C, c = O(1)$$



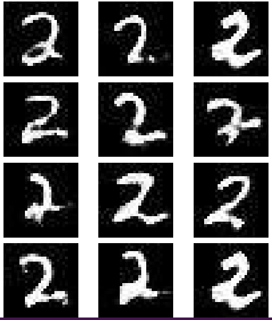


Table of contents

Regression in Machine Learning

Setting with 2 classes

- 2 laws in $\mathbb{R}^p$: $\mathcal{C}_+$; $\mathcal{C}_-$
- $X = (x_1, \dots, x_n) \in \mathcal{M}_{p,n}$: data matrix, $x_i \sim \mathcal{C}_+$ or $x_i \sim \mathcal{C}_-$
- notation: $\mu_a = \mathbb{E}[x_i]$, $\Sigma_a = \mathbb{E}[x_i x_i^T]$, for $x_i \sim \mathcal{C}_a$
- $Y \in \{-1, 1\}^n$: label vector $x_i \sim \mathcal{C}_a \Rightarrow y_i = a$

$\longrightarrow$ Look for $\beta \in \mathbb{R}^p$ s.t. $X^T\beta \approx Y$.

Ridge Regression

Minimise $\frac{1}{n}\left\|\beta^T X - Y\right\|^2 + \gamma\left\|\beta\right\|^2$, γ : regularising parameter

Robust Regression[3]

Minimise $\frac{1}{n}\sum_{i=1}^n f(y_i\beta^T x_i) + \gamma\|\beta\|^2$, for a loss function $f : \mathbb{R} \to \mathbb{R}$.

[3] El Karoui - 2013 : On robust regression with high-dimensional predictors

Table of contents

Ridge Regression

$$\text{Minimise} \quad \frac{1}{n}\sum_{i=1}^{n}(\beta^T x_i - y_i)^2 + \gamma \|\beta\|^2 .$$

Solution : $\beta = \frac{1}{n}QXY$ with $Q = \left(\frac{1}{n}XX^T + \gamma I_p\right)^{-1}$.

Performance estimation

- **Training** error: $E_{\mathrm{tr}} = \frac{1}{n}\|X^T\beta - Y\|^2$

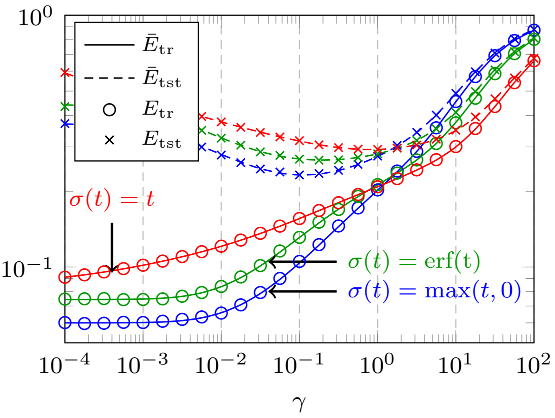
$$\bar{E}_{\mathrm{tr}} = \frac{1}{n}\mathbb{E}\left[\left\|\frac{1}{n}X^TQXY - Y\right\|^2\right] = f_1^\circ(\Sigma_\pm, \mu_\pm).$$

- **Test** error: $E_{\mathrm{tst}} = \frac{1}{n}\|X_t^T\beta - Y\|^2$, X_t, X i.i.d

$$\bar{E}_{\mathrm{tst}} = \frac{1}{n}\mathbb{E}\left[\frac{1}{n}YXQX_tX_t^TQXY - 2Y^TX_t^TQXY + Y^TY\right] = f_2^\circ(\Sigma_\pm, \mu_\pm$$

Example with One-Layer Neural Net $X = \sigma(WZ)$

- $Z = (z_1, \ldots z_n) \in \mathcal{M}_{q,n}$, MNIST data
- $W \in \mathcal{M}_{p,q}$, fixed initial drawing
- $\sigma : \mathbb{R} \to \mathbb{R}$: Lipschitz activation function



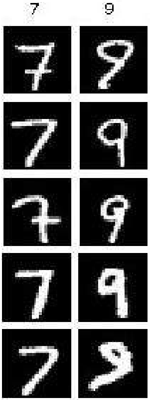


Table of contents

Robust Regression

$$\text{Minimise} \quad \frac{1}{n}\sum_{i=1}^{n} f(y_i x_i^T \beta) + \gamma\|\beta\|^2$$

Solution : $\beta = \frac{1}{n\gamma}\sum_{i=1}^{n} \phi(z_i^T \beta) z_i$ with $z_i = y_i x_i$ and $\phi = -\frac{1}{2} f'$

Theorem

If $X \propto \mathcal{E}_2$, $\phi : \mathbb{R} \to \mathbb{R}$ is λ-Lipschitz bounded and $\gamma > \frac{1}{\sqrt{n}}\lambda\|\mathbb{E}[X]\|^2$ then β is uniquely definedand:

$$\beta \propto \mathcal{E}_2\left(\frac{1}{\sqrt{n}}\right) \qquad \text{and} \qquad \mathbb{E}[\|\beta\|] = O(1).$$

$\to$ estimation of $\mathbb{E}[\beta]$ and $\mathbb{E}[\beta\beta^T]$ to predict performances

Statistics of β

Definition of $\beta_{-i}(t)$

$$\text{Sol}^\circ \text{ of } \beta_{-i}(t) = \frac{1}{n\gamma}\sum_{\substack{j=1\\ j\neq i}}^{n} \phi(z_j^T\beta_{-i}(t))z_j + \frac{1}{n}\underbrace{t\phi(z_i^T\beta_{-i}(t))}_{\chi_i(t)} z_i.$$

$\beta_{-i} = \beta_{-i}(0)$ and $\beta = \beta_{-i}(1)$.

1 - differentiation of $\beta_{-i}(t)$

$$\begin{aligned}\beta'_{-i}(t) &= \frac{1}{n\gamma}\sum_{\substack{j=1\\ j\neq i}}^{n} \underbrace{\phi'(z_j^T\beta_{-i}(t))}_{D_i(t)} z_j z_j^T \beta'_{-i}(t) + \frac{1}{n}\chi'_i(t) z_i\\ &= \frac{1}{n\gamma} Z_{-i} D(t) Z^T \beta'_{-i}(t) + \frac{1}{n}\chi'_i(t) z_i,\end{aligned}$$

with $Q(t) = \left(\frac{1}{n}Z_{-i}D(t)Z_{-i}^T - \frac{1}{\gamma}I_p\right)^{-1}$: $\beta'_{-i}(t) = \frac{1}{n}\chi'(t)Q(t)z_i$

Statistics of β

2 - Link between $z_i^T \beta_{-i}$ and $z_i^T \beta$

$$z_i^T \beta'_{-i}(t) = \chi'(t) \frac{1}{n} z_i^T Q(t) z_i$$

with $Q(t) = \left(\frac{1}{n} Z_{-i} D(t) Z_{-i}^T - \frac{1}{\gamma} I_p \right)^{-1}$ and $\chi(t) = t\phi(z_i^T \beta_{-i}(t))$

Proposition

$\left| \frac{1}{n} z_i^T Q(t) z_i - \frac{1}{n} z_i^T Q(0) z_i \right| = O(\frac{1}{\sqrt{n}})$and with $\delta \equiv \mathbb{E}[\frac{1}{n} z_i^T Q(0) z_i]$:

$$\frac{1}{n} z_i^T Q(t) z_i \in \delta \pm \mathcal{E}_2 \left(\frac{1}{\sqrt{n}} \right) + \mathcal{E}_1 \left(\frac{1}{n} \right)$$

Integration

$$\begin{aligned} z_i^T \beta - z_i^T \beta_{-i} &= \delta(\chi(1) - \chi(0)) + O\left(\frac{1}{\sqrt{n}} \right) \\ &= \delta \phi(z_i \beta) + O\left(\frac{1}{\sqrt{n}} \right) \end{aligned}$$

Statistics of β

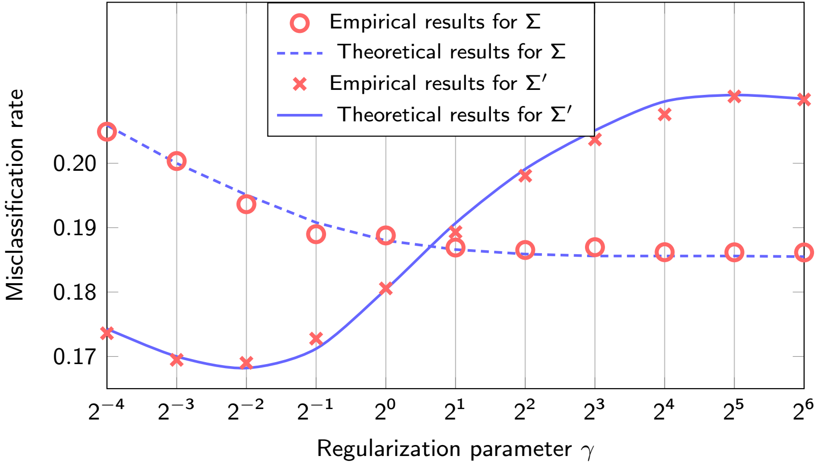
3 - computation of $\mathbb{E}[\beta]$ and $\mathbb{E}[\beta\beta^T]$

- β_{-i} and z_i independent $\Rightarrow z_i^T\beta_{-i} \sim \mathcal{N}(m_z^T m_\beta, \mathrm{Tr}(C_z C_\beta))$
- Deduce[4] m_β and Σ_β from:
  - $z_i^T\beta - z_i^T\beta_{-i} \approx \delta\phi(z_i^T\beta)$
  - $\beta = \frac{1}{n}\sum_{i=1}^n \phi(z_i^T\beta)z_i$

[4] Mai, Liao, Couillet - A Large Scale Analysis Of Logistic Regression: Asymptotic Performances And New Insights

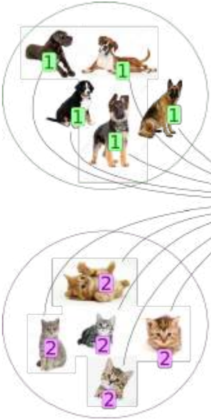
Application

- $p = 128,\ n = 512$
- $x_i \propto \mathcal{N}(y_i\mu, \Sigma)$
- $\Sigma = 2I_p$
- $\Sigma' = \mathrm{diag}[1,\ 5,\ \mathbf{1}_{p-2}]$

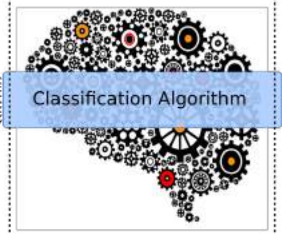


Conclusion

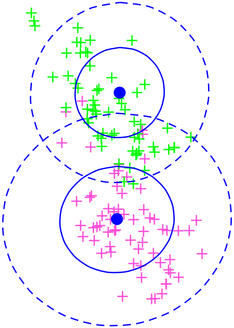
Real images
$\sim$ Concentrated vectors



Ridge Regression,
One-layer NN,
Robust Regression...



RMT + CMT tools
for Perf. Predictions



Conclusion

Random matrix theory allows for:

- precise estimate when $n = O(p)$

Combination with Concentration of Measure allows for

- extension to **realistic** hypotheses (GAN-generated data),
- tracking of concentration through **explicit** and **implicit** formulations,
- **rich** and **adaptable** characterisation of relevant quantities.

Thank you !

Table of contents

Position of the problem

Data matrix $X = (x_1, \ldots, x_n) \in \mathcal{M}_{p,n}$,

Hypothesis:

- $p = O(n)$ and $n = O(p)$
- $X \in \mathcal{E}_2(c)$
- $\|\mathbb{E}[X]\| = O(\sqrt{n})$

Goal:

Show the concentration of the resolvent:

$$Q = Q(z) = \left(\frac{1}{n}XX^T + zI_p\right)^{-1}$$

and find a computable *deterministic equivalent* $\tilde{Q}_1$ depending on the population covariance : $\Sigma = \frac{1}{n}\mathbb{E}[XX^T]$

Basic results on the resolvent $Q = \left(\frac{1}{n}XX^T + zI_p\right)^{-1}$

- The resolvent is bounded:

$$\|Q(z)\| \leq \frac{1}{z}, \left\| Q(z)\frac{XX^T}{n} \right\| \leq 1 \text{ and } \left\| Q(z)\frac{X}{\sqrt{n}} \right\| \leq \frac{1}{z^{1/2}}$$

- $X \mapsto Q(z)$ is $\frac{1}{\sqrt{n}z^{3/2}}$-Lipschitz:

If we note $Q(z)^H = \left(\frac{1}{n}(X+H)(X+H)^T + zI_p\right)^{-1}$:

$$\begin{aligned}
\left\| Q(z)^H - Q(z) \right\|_F &= \left\| \frac{1}{n} Q(z)^H (XX^T - (X+H)(X+H)^T) Q(z) \right\|_F \\
&= \left\| -\frac{1}{n} Q(z)^H HX^T + (X+H)H^T) Q(z) \right\|_F \\
&\leq \frac{1}{\sqrt{n}} \left(\|Q(z)^H\| \left\| \frac{1}{\sqrt{n}} X^T Q \right\| + \left\| \frac{1}{\sqrt{n}} Q^H (X+H) \right\| \|Q(z)\| \right) \|H\|_F
\end{aligned}$$

- $Q(z) \in \mathbb{E}[Q(z)] \pm C\mathcal{E}_2\left(\frac{c}{\sqrt{n}}\right)$ (we suppose that $\frac{1}{z} = O(1)$)

Question

How to estimate $\mathbb{E}\left[\left(\frac{1}{n}XX^T + zI_p\right)^{-1}\right]$ *?*

Design of a Deterministic equivalent

Let $\tilde{\Sigma} \in \mathcal{M}_p$ to be chosen precisely later and we set:

$$\boxed{\tilde{Q}_1 = (\tilde{\Sigma} + zI_p)^{-1}}$$

With identity $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$:

$$\mathbb{E}[\tilde{Q}_1 - Q] = \mathbb{E}\left[Q\left(\frac{1}{n}XX^T - \tilde{\Sigma}\right)\tilde{Q}_1\right] \quad = \sum_{i=1}^n \frac{1}{n}\mathbb{E}\left[Q(x_i x_i^T - \tilde{\Sigma})\tilde{Q}_1\right].$$

Schur formulas

We set $X_{-i} = (x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n) \in \mathcal{M}_{p,n}$
and $Q_{-i} = (\frac{1}{n}X_{-i}X_{-i}^T + zI_p)^{-1}$:

$$Q = Q_{-i} - \frac{1}{n}\frac{Q_{-i}x_i x_i^T Q_{-i}}{1 + \frac{1}{n}x_i^T Q_{-i}x_i} \quad \text{and} \quad Qx_i = \frac{Q_{-i}x_i}{1 + \frac{1}{n}x_i^T Q_{-i}x_i}.$$

Then:

$$\tilde{Q}_1 - \mathbb{E}Q = \sum_{i=1}^n \frac{1}{n}\mathbb{E}\left[Q_{-i}\left(\frac{x_i x_i^T}{1 + \frac{1}{n}x_i^T Q_{-i}x_i} - \tilde{\Sigma}\right)\tilde{Q}_1\right]$$
$$- \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[Q_{-i}x_i x_i^T Q\tilde{\Sigma}\tilde{Q}_1\right].$$

A first deterministic equivalent

$$\left\|\tilde{Q}_1 - \mathbb{E}Q\right\| = \sup_{\|u\|,\|v\|\leq 1} u^T\left(\tilde{Q}_1 - \mathbb{E}Q\right)v$$

$$= \sup_{\|u\|,\|v\|\leq 1} \frac{1}{n}\sum_{i=1}^{n} \Delta_i + \varepsilon_i$$

with:

- $\Delta_i = \mathbb{E}\left[u^T Q_{-i}\left(\frac{x_i x_i^T}{1+\frac{1}{n}x_i^T Q_{-i}x_i} - \tilde{\Sigma}\right)\tilde{Q}_1 v\right]$
- $\varepsilon_i = \frac{1}{n}\mathbb{E}\left[u^T Q_{-i}x_i x_i^T Q\tilde{\Sigma}\tilde{Q}_1 v\right]$

$\longrightarrow$ we note $\delta_1 = \frac{1}{n}\mathrm{Tr}(\Sigma\mathbb{E}[Q_{-i}])$ and we chose $\boxed{\tilde{\Sigma} = \frac{\Sigma}{1+\delta_1}}$

Let us show that with this choice: $\Delta_i, \varepsilon_i = O\left(\frac{1}{\sqrt{n}}\right)$

Preliminary lemmas

- $u^T Q x_i = \frac{1}{\sqrt{n}}(\sqrt{n} u^T Q) x_i \in \mathcal{E}_2(c) + C\mathcal{E}_1\left(\frac{c}{\sqrt{p}}\right)$

- $$\mathbb{E}[u^T Q x_i] \leq \sqrt{\mathbb{E}[u^T Q x_i x_i^T Q u]} = \sqrt{\frac{1}{n}\mathbb{E}[u^T Q X X^T Q u]}.$$
$$\leq \mathbb{E}\left[\left\|u^T Q u\right\|\right] = O(1)$$

- The same way:
$u^T Q_{-i} x_i, u^T \tilde{Q}_1 x_i \in O(1) \pm C\mathcal{E}_2(c) + C\mathcal{E}_1\left(\frac{c}{\sqrt{p}}\right)$

Preliminary lemmas

- $\frac{1}{n}x_i^T Q_{-i} x_i \in \mathcal{E}_2(c) + C\mathcal{E}_1\left(\frac{c}{\sqrt{n}}\right)$
- $\mathbb{E}\left[\frac{1}{n}x_i^T Q_{-i} x_i\right] = \frac{1}{n}\operatorname{Tr}(\Sigma\mathbb{E}[Q_{-i}]) \leq \frac{1}{n}\operatorname{Tr}(\Sigma)\mathbb{E}\left[\|Q_{-i}\|\right] = O(1)$
-

$$\|\mathbb{E}Q_{-i} - \mathbb{E}Q\| = \sup_{\|u\|,\|v\|\leq 1} u^T\left(\mathbb{E}Q_{-i} - \mathbb{E}Q\right)v$$

$$= \sup_{\|u\|,\|v\|\leq 1} \mathbb{E}\left[\frac{1}{n}u^T Q_{-i} x_i x_i^T Q v\right] = \sup_{\|u\|,\|v\|\leq 1} \frac{1}{n}\sqrt{\mathbb{E}[}$$

End of the proof, $\tilde{Q}_1 = \left(\frac{\Sigma}{1+\delta_1} + zI_p\right)^{-1}$

- Since $\left\|\tilde{\Sigma}\tilde{Q}_1\right\| = O(1)$, with Holder inequality :

$$\varepsilon_i = \frac{1}{n}\mathbb{E}\left[u^T Q_{-i} x_i x_i^T Q \tilde{\Sigma} \tilde{Q}_1 v\right]$$

$$\leq \frac{1}{n}\sqrt{\mathbb{E}\left[(u^T Q_{-i} x_i)^2\right]\mathbb{E}\left[(x_i^T Q \tilde{\Sigma}\tilde{Q}_1 v)^2\right]} = O\left(\frac{1}{n}\right)$$

- $$\Delta_i = \mathbb{E}\left[u^T Q_{-i}\left(\frac{x_i x_i^T}{1 + \frac{1}{n}x_i^T Q_{-i} x_i} - \frac{\Sigma}{1+\delta_1}\right)\tilde{Q}_1 v\right]$$

$$= \mathbb{E}\left[\frac{u^T Q_{-i} x_i x_i^T \tilde{Q}_1 v(\delta_1 - \frac{1}{n}x_i^T Q_{-i} x_i)}{\left(1 + \frac{1}{n}x_i^T Q_{-i} x_i\right)(1+\delta_1)}\right]$$

$$+ \mathbb{E}\left[u^T Q_{-i}\left(\frac{x_i x_i^T - \Sigma}{1+\delta_1}\right)\tilde{Q}_1 v\right] = O\left(\frac{1}{\sqrt{n}}\right)$$

$$\Longrightarrow \left\|\mathbb{E}[Q] - \tilde{Q}_1\right\| = O\left(\frac{1}{\sqrt{n}}\right)$$

Second deterministic equivalent

Note that $\delta_1 = \frac{1}{n}\operatorname{Tr}(\Sigma\mathbb{E}[Q]) = \frac{1}{n}\operatorname{Tr}(\Sigma\tilde{Q}_1) + O\left(\frac{1}{\sqrt{n}}\right)$

$$= \frac{1}{n}\operatorname{Tr}\left(\Sigma\left(\frac{\Sigma}{1+\delta_1} + zI_p\right)^{-1}\right) + O\left(\frac{1}{\sqrt{n}}\right)$$

The function

$$\begin{array}{ccc} \mathbb{R}^+ & \longrightarrow & \mathbb{R}^+ \\ \delta & \longmapsto & \frac{1}{n}\operatorname{Tr}\left(\Sigma\left(\frac{\Sigma}{1+\delta} + zI_p\right)^{-1}\right) \end{array}$$

is contracting for the semimetric: $d_s(\delta,\delta') = \frac{|\delta-\delta'|}{\sqrt{\delta\delta'}}$

$\Longrightarrow$ It admits a unique fixed point:

$$\delta_2 = \frac{1}{n}\operatorname{Tr}\left(\Sigma\left(\frac{\Sigma}{1+\delta_2} + zI_p\right)^{-1}\right)$$

End of the proof

It can be showed that $\delta_1 - \delta_2 = O\left(\frac{1}{\sqrt{n}}\right)$ thus if we set

$$\boxed{\tilde{Q}_2 = \left(\frac{\Sigma}{1+\delta_2} + zI_p\right)^{-1}}:$$

$$\left\|\mathbb{E}[Q] - \tilde{Q}_2\right\| \leq \left\|\mathbb{E}[Q] - \tilde{Q}_1\right\| + \left\|\tilde{Q}_1 - \tilde{Q}_2\right\|$$

$$\leq O\left(\frac{1}{\sqrt{n}}\right) + \left\|\tilde{Q}_1 \frac{\Sigma(\delta_2 - \delta_1)}{(1+\delta_2)(1+\delta_1)}\tilde{Q}_2\right\| \; = \; O\left(\frac{1}{\sqrt{n}}\right)$$

$$\boxed{\begin{array}{l} \Longrightarrow Q \in \tilde{Q}_1 \pm C\mathcal{E}_2\left(\frac{c}{\sqrt{n}}\right) \\ \Longrightarrow \forall t > 0 \; : \; \mathbb{P}\left(\left|\frac{1}{p}\operatorname{Tr}(AQ) - \frac{1}{p}\operatorname{Tr}(A\tilde{Q}_2)\right| \geq t\right) \leq Ce^{-cnt^2} \end{array}},$$

(for $A \in \mathcal{M}_p$, $\|A\|_1 \leq p$)

Table of contents

Table of contents