Apprentissage Statistique en Grande Dimension

F. Panloup
LAREMA-Université d'Angers

—

L'EFFET DE LA DIMENSION

—

October 3, 2019

Qu'est-ce que la grande dimension ? Cadres pratiques

Exemples

1. Données biotech: mesure des milliers de quantités par "individu". On peut penser en particulier aux données "omiques" où l'on cherche à appréhender une maladie par exemple via un ensemble d'information par individu de l'ordre de plusieurs dizaines à centaines de milliers de variables (cf schéma ci-dessous issu de `http://www.ipubli.inserm.fr/`)

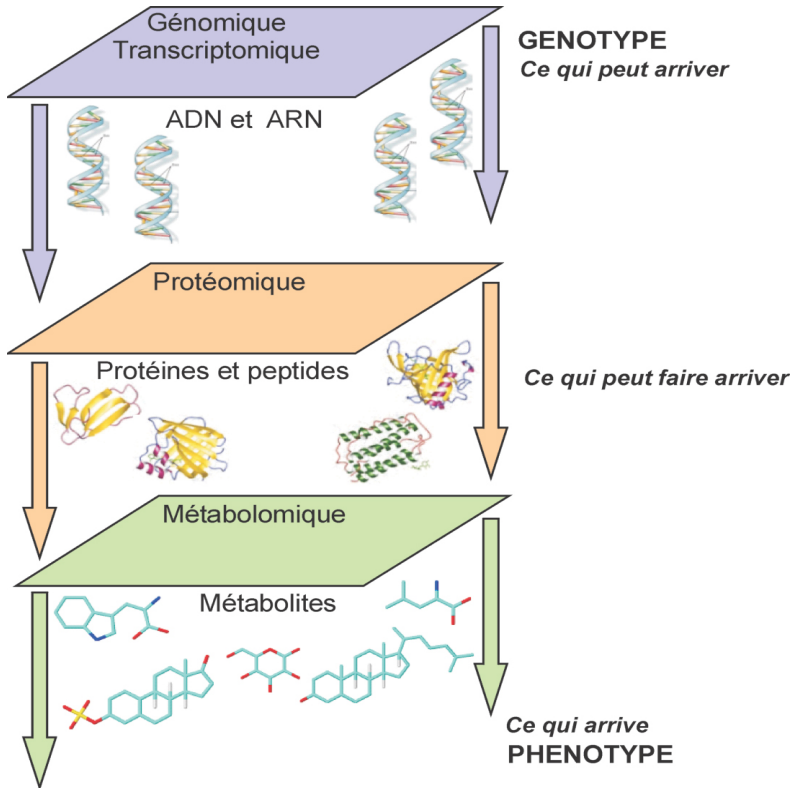


Figure 1: Données omiques

Exemples

1. Traitement d'images: images médicales, astrophysique, video surveillance, etc. Chaque image est constituée de milliers ou millions de pixels ou voxels.
2. Marketing : les sites web et les programmes de fidélité collectent de grandes quantités d'information sur les préférences et comportements des clients. Ex: syst'emes de recommandation...
3. Business : exploitation des données internes et externes de l'entreprise devient primordial

n et p

- Dans tous ces champs d'applications, l'objectif est de décrire de manière plus précise un phénomène (penser au traitement d'images ou à la médecine par exemple). Néanmoins, sous quelles conditions l'accumulation de données peut permettre d'aller dans cette direction ?
- Quantités fondamentales : n le nombre d'observations/individus (patients/images/"clics"...) et p, le nombre de variables/mesures par "individu" (éviter le terme "données" qui peut porter à confusion).
- On parle de Grande Dimension lorsque p est grand, *i.e.* lorsque l'espace sur lequel "vivent" les observations est de grande dimension.
- Du point de vue informatique, le caractère n grand est lui AUSSI un facteur de "Big Data" (puisque le temps de calcul s'en trouve impacté) mais on comprendra que du point de vue mathématique, "plus n est grand, plus on est content".
- Plusieurs situations : en médecine, $n << p$, en traitement d'images récoltées sur Internet (cf Facebook par exemple), on peut imaginer n et p grands.
- Ces deux types de situations sont totalement différents. En particulier, les objectifs d'apprentissage sont totalement dépendants du lien entre ces deux valeurs.

De la statistique classique à la statistique en grande dimension

n, p et statistique

1. Statistique Classique : p petit et n grand. Ainsi, on peut étudier le problème de manière asymptotique via les résultats classiques type TCL par exemple : $f : \mathbb{R}^p \to \mathbb{R}$, $X_1, \dots, X_n$ $i.i.d.$

$$\sqrt{n}\frac{\left(\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_1)]\right)}{\sqrt{\mathrm{Var}(f(X_1))}} \xrightarrow{n\to+\infty} \mathcal{N}(0,1).$$

2. Le problème est que $\mathrm{Var}(f(X_1))$ augmente avec p. Supposons que f soit Lipschitzienne. Dans ce cas, on a

$$\mathrm{Var}(f(X_1)) \leq \mathbb{E}[(f(X_1) - \mathbb{E}[f(X_1)])^2] \leq C\mathbb{E}[\|X_1\|^2] = C\sum_{i=1}^p \mathbb{E}[(X_1^j)^2] \propto p$$

   si les coordonnées ont la même loi par exemple.

3. Dans un cadre général, le ratio $\sqrt{\frac{p}{n}}$ est donc fondamental dans le calcul de l'erreur.

n, p et statistique

- Statistique en grande dimension : $n << p$ ou $n \propto p$. De manière générale, il faut repenser la statistique dans un mode non asymptotique en n (Chebyshev/Concentration) mais aussi et surtout adapter les objectifs.
- Par exemple, travailler sur un sous-espace/une sous-variété de $\mathbb{R}^p$ qui soit de dimension adaptée au nombre d'observations (LASSO).
- Utiliser des méthodes de classification qui "supportent la dimension" (arbres/SVM...)
- Réduire la dimension du problème...

Curse of Dimensionality

Grande Dimension et Méthodes à moyennage local

On a vu dans le chapitre 1 que les méthodes à moyennage local (type KNN) sont des candidats non paramétriques naturels pour l'apprentissage. Qu'en est-il en grande dimension ?

- **Exemple** : Supposons que l'on ait à apprendre une relation de la forme $Y = f(\mathbf{X}, \varepsilon)$ où $\mathbf{X}$ suit la loi uniforme sur l'hypercube $[0,1]^p$.
- Supposons même pour simplifier que $Y = f(\mathbf{X})$ (*i.e.* que sachant $\mathbf{X}$ la réponse est déterministe). Supposons même pour simplifier encore plus que l'on subdivise chaque dimension en 10 (sur chaque dim., on divise l'intervalle $[0,1]$ en 10 intervalles) et que f est constante sur les hypercubes de la forme $\prod_{k=1}^{p} [i_k/10, i_{k+1}/10]$. Dans ce cas, le nombre minimal d'observations pour apprendre la relation $Y = f(\mathbf{X})$ est égal à

  $$10^p !! \quad \text{(Croissance exponentielle avec la dimension).}$$

- En langage plus élaboré, cette propriété peut s'interpréter comme la décroissance exponentielle du volume de la boule unité avec la dimension.

Boule unité en grande dimension

On a :

$$V_p(1) = \mathrm{Vol}(B_p(0,1)) = \frac{\pi^{\frac{p}{2}}}{\Gamma(\frac{p}{2}+1)} \overset{p\to+\infty}{\sim} \left(\frac{2\pi e}{p}\right)^{\frac{p}{2}} (p\pi)^{-\frac{1}{2}}.$$

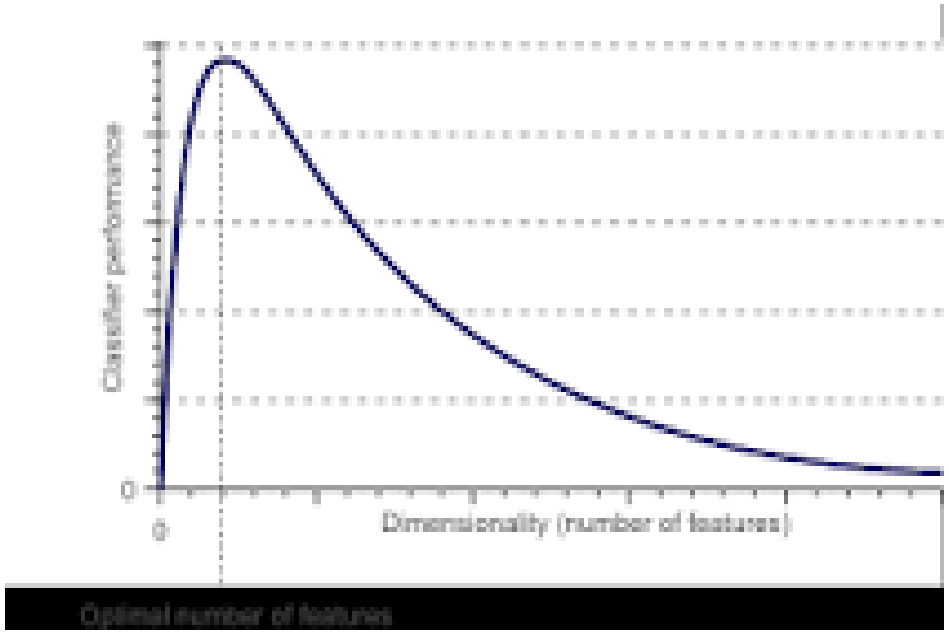


Figure 2: $p \to \mathrm{Vol}(B_p(0,1))$

Boule unité en grande dimension et nombre d'observations

Ainsi, si $x^{(1)}, \dots x^{(n)}$ sont n points tels que pour tout $x \in [0,1]^p$, il existe au moins un point parmi les n tel que $\|x^{(i)} - x\| \leq 1$, on a alors

$$[0,1]^p \subset \cap_{i=1}^n B(x^{(i)}, p)$$

de sorte que $1 \leq nV_p(1)$. On retrouve que le nombre de points nécessaires pour remplir l'hypercube satisfait :

$$n \geq \frac{\Gamma(\frac{p}{2}+1)}{\pi^{\frac{p}{2}}} \sim \left(\frac{p}{2\pi e}\right)^{\frac{p}{2}} (p\pi)^{\frac{1}{2}}$$

En dimension 100, on trouve déjà $n \geq 42.10^{39}$.

Répartition de la masse d'une boule en grande dimension

Dans le même sens, on peut remarquer que la boule unité en grande dimension peut s'avérer peu intuitive. Considérons la boule de rayon r et la couronne $C_p(r) = \{x, 0.99r \leq \|x\| \leq r\}$. On a ‘

$$\frac{\mathrm{Vol}(C_p(r))}{\mathrm{Vol}(B_p(r))} = 1 - 0.99^p \to 1$$

lorsque $p \to +\infty$ (exponentiellement vite). Ainsi, très rapidement, la masse de la boule unité se trouve concentrée dans sa “croûte”. La répartition des points dans l'espace est finalement assez surprenante.

Moyennage local ?

1. Au vu de ce qui précède, on ne peut clairement pas envisager d'utiliser des méthodes à moyennage local en grande dimension.
2. Utiliser les k-plus proches voisins n'a donc pas de sens en général car il n'y a plus de voisins en grande dimension.
3. Ceci est confirmé par la borne théorique obtenue par Gadat *et. al.*:

$$\sup_{(\mathbf{X},Y)} |\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^\star)| \leq C n^{-\frac{1+\alpha}{2+d}}.$$

   où $\mathcal{R}(f) = \mathbb{E}[\ell(Y, f(\mathbf{X}))]$.
4. Néanmoins, la borne ci-dessus peut être vue comme pessimiste car elle est "universelle" (pour l'ensemble des lois $(\mathbf{X}, Y)$ telles que $x \mapsto \eta(x) := \mathbb{P}(Y = 1 | \mathbf{X} = x)$ varie raisonnablement par exemple).
5. *Cuisine :* Si la fonction η est très peu sensible ou si le support "réel" de X est de dimension plus faible, cela peut encore fonctionner.

Exemple de la régression linéaire

Supposons qu'il existe $\theta^\star \in \mathbb{R}^p$ tel que pour tout $i \in \{1, \dots, n\}$,

$$Y_i = \langle \mathbf{x}_i, \theta^\star \rangle + \varepsilon_i$$

où $Y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^p$ et $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Posons $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^T$. L'estimateur des moindres carrés $\hat{\theta}$ satisfait (lorsque $\mathbf{x}^T\mathbf{x}$ est inversible):

$$\hat{\theta} = \mathrm{Argmin}_{\theta \in \mathbb{R}^p} \| Y - \mathbf{x}\theta \|^2 = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T Y.$$

De plus, si $\mathbf{x}^T\mathbf{x}$ est inversible (matrice $n \times p$),

$$\mathbb{E}[\|\hat{\theta} - \theta_0\|^2] = \mathbb{E}[\|(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\varepsilon\|^2] = \mathrm{Tr}((\mathbf{x}^T\mathbf{x})^{-1})\sigma^2.$$

Supposons que les colonnes $C_1, \dots, C_p$ de $\mathbf{x}$ soient orthonormales (ce qui implique que $p \leq n$). Dans ce cas,

$$(\mathbf{x}^T\mathbf{x})_{i,j} = \langle C_i, C_j \rangle = \delta_{i,j} \implies \mathbf{x}^T\mathbf{x} = I_p.$$

Par conséquent,

$$\mathbb{E}[\|\hat{\theta} - \theta_0\|^2] = p\sigma^2.$$

On retrouve à nouveau la croissance linéaire de l'erreur de la MSE avec la dimension.

Matrice de covariance

Soit X une variable aléatoire à valeurs dans $\mathbb{R}^p$ de matrice de covariance Σ (matrice $p \times p$). Pour simplifier, supposons que X est centrée. Dans ce cas, $\Sigma_{i,j} = \mathbb{E}[X_i X_j]$ et notons $(X^{(1)}, \dots, X^{(n)})$ un échantillon $i.i.d.$ issu de X.
Soit $\Sigma^{(n)}$ la matrice de covariance empirique associée :

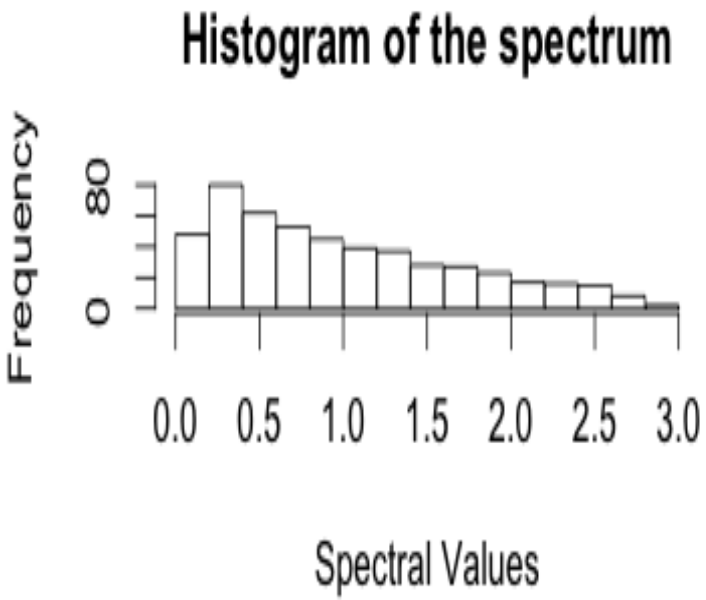
$$\Sigma^{(n)}_{i,j} = \frac{1}{n} \sum_{k=1}^{n} X_i^{(k)} X_j^{(k)}.$$

Par la loi des grands nombres, si p est fixé et $n \to +\infty$,

$$\Sigma^{(n)} \to \Sigma.$$

Question : Si p n'est pas petit devant n, peut-on encore espérer que $\Sigma^{(n)}$ soit une bonne approximation de Σ ?
Réponse : Non, en général. Par exemple, on peut le voir sur le spectre de la matrice de covariance empirique. Dans les graphes suivants, on suppose que $X \sim \mathcal{N}(0, I_p)$ et on remarque que le spectre de la matrice de covariance ne ressemble pas du tout au spectre de la matrice cible.

Matrice de covariance (exemple)



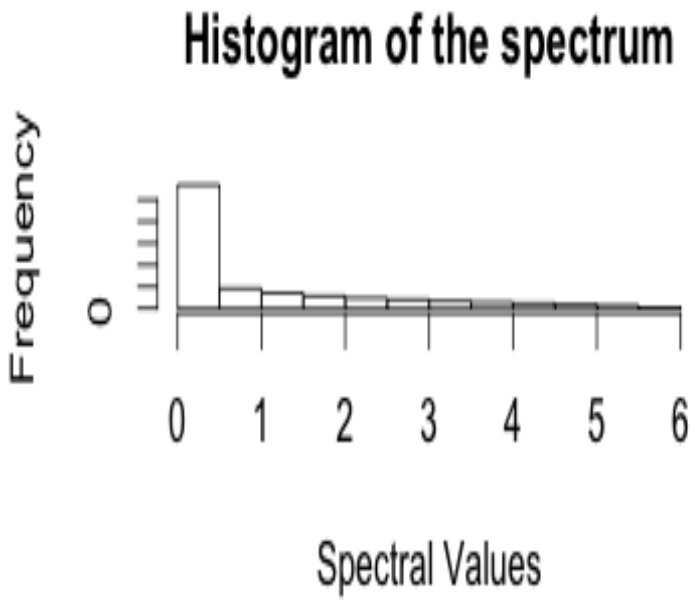


Figure 3: $n = 1000$, $p = 500$ (left), $p = 2000$ (right)

Et l'ACP ?

On rappelle que l'Analyse en Composantes Principales consiste pour un d fixé à déterminer le meilleur sous-espace vectoriel V_d de dimension d au sens suivant :

$$V_d^{(n)} \in \mathrm{Argmin}_{dim V \leq d} \sum_{i=1}^{n} \|X^{(i)} - P_V(X^{(i)})\|^2$$

avec l'objectif sous-jacent de déterminer le meilleur sous-espace vectoriel au sens suivant :

$$V_d \in \mathrm{Argmin}_{dim V \leq d} \mathbb{E}[\|X - P_V(X)\|^2].$$

Via la loi des grands nombres, on a "$V_d^{(n)} \to V_d$" lorsque $n \to +\infty$.
Question : L'intérêt de l'ACP est ainsi de réduire la dimension de l'espace "optimalement" (en perdant le moins d'information) mais est-ce que cette réduction de dimension est fiable ? Sous quelles conditions ?

Et l'ACP ? (suite)

On rappelle le résultat suivant (lorsque les variables sont centrées, sinon il faut recentrer).

Théorème

Supposons que l'échantillon soit issu d'une v.a. X centrée de matrice de covariance $\mathbb{E}[XX^T]$ de rang supérieur à d. Dans ce cas, $V_d^{(n)}$ est le sous-espace vectoriel de dimension d engendré par les d vecteurs propres associés aux plus grandes valeurs propres de $\Sigma^{(n)}$, la matrice de covariance empirique définie par :

$$\Sigma_{i,j}^{(n)} = \frac{1}{n} \sum_{k=1}^{n} X_i^{(k)} X_j^{(k)}.$$

Pour rappel, les matrices de covariance (empirique ou non) sont symétriques positives. Ainsi, quitte à normaliser, les vecteurs propres forment une base orthonormée de V_d.

Au vu de ce qu'on a dit sur la matrice de covariance précédemment, on peut donc en conclure l'ACP ne donne pas d'information fiable en grande dimension et en toute généralité que sous la condition $n >> p$.

Néanmoins, si la matrice de covariance est de rang faible ou contient un grand nombre de petites valeurs propres, l'ACP peut s'avérer encore efficace. A voir en

Que peut-on espérer en grande dimension ?

Les messages ci-dessous (qui ne sont que des exemples de difficultés rencontrées) ne sont pas très rassurants pour aborder des problèmes en grande dimension. Néanmoins, comme l'indique la suggestion dans le cadre de l'ACP pour les matrices de faible rang, on peut espérer tirer de l'information si :

- il existe des structures "cachées" dans les espaces de grande dimension qui sont de petite dimension: les données en grande dimension sont concentrées autour de structures de faible dimension reflétant la faible complexité des données. On peut penser à la structure géométrique des images par exemple. En médecine, ce type d'hypothèse n'est pas clair pour certaines maladies complexes mais le faible nombre de données nous astreint à une telle hypothèse.
- La structure est plus complexe mais le nombre de données est lui aussi élevé ce qui permet via des algorithmes performants d'aller plus loin dans la flexibilité et donc d'augmenter la prédiction, l'analyse...