Statistiques en grande dimension

Léo Gayral

Ces notes sont basées sur le cours de Christophe Giraud.

Table des matières

1 Sélection de modèle

1.1 Modèles et oracles

On s'intéresse ici à des modèles de régression linéaire, comme définis ci-dessous.

Définition 1 (Modèle linéaire) **:**
On considère n variables réelles $Y_i = \langle X(i), \beta^* \rangle + \varepsilon_i \in \mathbb{R}$, où chaque vecteur $X^{(i)} \in \mathbb{R}^p$ contient les variables explicatives de Y_i, et les variables $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ représentent le bruit.

On peut réécrire cette équation sous la forme matricielle $Y = X\beta^* + \varepsilon \in \mathbb{R}^n$, et on notera parfois $f^* := X\beta^*$.

Remarque 2 (Structure supplémentaire) **:**
On considèrera parfois $\|\beta^*\|_0 \ll p$, autrement dit une hypothèse de parcimonie sur le modèle.

On pourra étendre cette notion à une parcimonie par paquets. Plus précisément, si on partitionne $[p] = \bigsqcup_{k=1}^{M} G_k$, on voudra $\#\{k,\ \beta^*|_{G_k} \neq 0\} \ll M$.

Par la suite, l'objectif sera de prédire au mieux la valeur de β^* ou de f^*.

1.1.1 Structure connue

Supposons ici qu'on connait $m^* = \{k,\ \beta_k^* \neq 0\}$. Autrement dit, on sait quelles variables jouent un rôle sur Y. Dans ce cas, on peut faire une simple régression linéaire en $|m^*|$ dimensions.

Plus largement, si on cherche f^* dans un espace S, alors on peut considérer le maximum de vraisemblance $\widehat{f_S} = \underset{f \in S}{\operatorname{argmax}}\, L(f)$. Lorsque S est un sous-espace vectoriel, $\widehat{f_S}$ est simplement la projection orthogonale de Y sur ce sous-espace.

1.1.2 Structure inconnue

Dans ce cas, on doit comparer une famille de modèles, d'espaces $(S_m)_{m \in \mathcal{M}}$ dans lesquels f^* pourrait se trouver. On calcule alors chaque $\widehat{f_m} = p_{S_m}(Y)$, puis on choisit la meilleure prédiction pour une certaine métrique. Par la suite, on minimisera typiquement le risque définie ci-dessous.

Dans le cas de la parcimonie par paquets, on considère typiquement $\mathcal{M} = \mathcal{P}([M])$ avec les modèles $S_m = \bigcup_{k \in m} G_k$ correspondant au choix des paquets de coordonnées de β à conserver.

Définition 3 (Risque quadratique) **:**
On définit $R(f) = \mathbb{E}\big[\|f - f^*\|_2^2\big]$.

Lemme 4 :
Soient $Z \in \mathbb{R}^d$ une variable aléatoire et $A \in M_d(\mathbb{R})$ une matrice. Alors :
— $\mathrm{Cov}(AZ) = A\mathrm{Cov}(Z)A^T$,
— $\mathbb{E}\big[\|Z\|^2\big] = \|\mathbb{E}[Z]\|^2 + \mathrm{Tr}(\mathrm{Cov}(Z))$.

Lemme 5 :
Si $\widehat{f}_S = p_S(Y)$ est une projection orthogonale, alors le risque quadratique vérifie :

$$R\Big(\widehat{f}_S\Big) = \mathbb{E}\big[\|p_S(\varepsilon)\|^2\big] + \|f^* - p_S(f^*)\|^2 = \sigma^2 \dim(S) + \|f^* - p_S(f^*)\|^2 .$$

On notera $d_m = \dim(S_m)$ par la suite.

Définition 6 (Oracle) **:**
Un estimateur oracle $\widehat{f}_{m_0}$ est un estimateur qui minimise le risque pour $m \in \mathcal{M}$.

En pratique, on ne peut pas calculer $R(f)$ car il faudrait connaître f^* au préalable. On ne peut donc pas calculer l'oracle. On va donc chercher à obtenir des risques empiriques $\widehat{R}\Big(\widehat{f}_m\Big)$ et prendre l'estimateur $\widehat{f}_{\widehat{m}}$ qui minimise le risque empirique.

Se posent alors les questions de comment calculer $\widehat{R}$, de la valeur de $R\Big(\widehat{f}_{\widehat{m}}\Big)$, et de l'éventuelle optimalité de cet estimateur.

1.2 Approche historique

L'approche naïve est de considérer $\widehat{R}(f) = \|Y - f\|^2$. Cette approche peut fonctionner lorsqu'on considère des modèles de taille fixée, mais pour des modèles emboités $S \subset T$, on aura toujours $\widehat{R}\Big(\widehat{f}_T\Big) \leq \widehat{R}\Big(\widehat{f}_S\Big)$.

Un calcul explicite donne $\mathbb{E}\Big[\widehat{R}\Big(\widehat{f}_m\Big)\Big] = \sigma^2(n - d_m) + \|f^* - p_S(f^*)\|^2 = R\Big(\widehat{f}_m\Big) + (n - 2d_m)\sigma^2$. Cet estimateur est donc biaisé. En admettant qu'on connait la variance du bruit σ, on peut alors définir un estimateur de risque sans biais.

Définition 7 (Risque AIC) **:**
On définit $\widehat{R}_{AIC}\Big(\widehat{f}_m\Big) = \|Y - \widehat{f}_m\|^2 + (2d_m - n)\sigma^2$.

Cette approche fonctionne relativement bien pour $|\mathcal{M}|$ faible, mais échoue lorsqu'on considère des grandes familles de modèles.

2 Bornes sur le risque

2.1 Majorations et critères optimisés

Le but est ici de choisir un bon candidat pour $\widehat{R}$. Par la suite, on considère des candidats sous la forme $\widehat{R}\left(\widehat{f}_m\right) = \left\|Y - \widehat{f}_m\right\|^2 + \sigma^2 \text{pen}(m)$, où $pen : \mathcal{M} \to \mathbb{R}^+$ est la fonction de pénalité.

Pour mieux choisir la fonction de pénalité pen, on veut désormais obtenir une majoration du type $R\left(\widehat{f}_{\widehat{m}}\right) \leq C + C' \min_{m \in \mathcal{M}} R\left(\widehat{f}_m\right)$. On appelle ce genre de résultats des inégalités d'oracle.

Lemme 8 (Inégalité de concentration gaussienne) **:**

Pour toute fonction F 1-lipschitzienne sur $\mathbb{R}^n$ et tout vecteur $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, il existe une variable exponentielle $\xi_F \sim \mathcal{E}(1)$ telle que $F(\varepsilon) \leq \mathbb{E}[F(\varepsilon)] + \sigma\sqrt{2\xi_F}$.

Définition 9 :

Soit $\xi \sim \mathcal{E}(1)$ une variable exponentielle. On définit l'application suivante :

$$R(d, \alpha) = \mathbb{E}\left[\left(\left(\sqrt{d} + \sqrt{2\xi}\right)^2 - \alpha\right)^+\right].$$

Théorème 10 :

Soient un réel $a > 1$ et $\mathcal{M}$ une famille de modèles. On définit $\rho_a(\mathcal{M}) = 1 + \sum_{m \in \mathcal{M}} R\left(d_m, \frac{\text{pen}(m)}{a}\right)$. On a alors la majoration :

$$\frac{a-1}{a} R\left(\widehat{f}_{\widehat{m}}\right) \leq \inf_{m \in \mathcal{M}} \left(R\left(\widehat{f}_m\right) + \sigma^2 \text{pen}(m)\right) + a\sigma^2 \rho_a(\mathcal{M}).$$

Démonstration. Étant donné le inf, il suffit de montrer l'inégalité pour $m \in \mathcal{M}$ fixé.

Par définition, $\widehat{R}\left(\widehat{f}_{\widehat{m}}\right) \leq \widehat{R}\left(\widehat{f}_m\right)$. En passant à l'espérance et en simplifiant via $Y = f^* + \varepsilon$, on obtient alors :

$$R\left(\widehat{f}_{\widehat{m}}\right) \leq R\left(\widehat{f}_m\right) + \sigma^2 \text{pen}(m) + \mathbb{E}\left[2\left\langle \varepsilon, \widehat{f}_{\widehat{m}} - f^* \right\rangle - \sigma^2 \text{pen}(\widehat{m})\right].$$

On cherche à obtenir $2\left\langle \varepsilon, \widehat{f}_{\widehat{m}} - f^* \right\rangle - \sigma^2 \text{pen}(\widehat{m}) \leq \frac{1}{a}\left\|\widehat{f}_{\widehat{m}} - f^*\right\|^2 + a\sigma^2 Z$, ainsi que la majoration en espérance $\mathbb{E}[Z] \leq \rho_a(\mathcal{M})$. On utilisera pour cela l'inégalité $2\langle x, y\rangle \leq \frac{1}{a}\|x\|^2 + a\|y\|^2$.

Posons $\overline{S_m} := \text{Vect}(f^*, S_m) = \mathbb{R}f^* \overset{\perp}{\oplus} \widetilde{S_m}$. Comme $\widehat{f}_{\widehat{m}} - f^* \in \overline{S_{\widehat{m}}}$, on peut se ramener à un produit scalaire avec $p_{\overline{S_m}}(\varepsilon)$. Définissons le carré de gaussienne $N^2 := \frac{\left\|p_{\mathbb{R}f^*}(\varepsilon)\right\|^2}{\sigma^2} \sim \mathcal{N}(0,1)^2$, et

$U_m := \frac{\left\|p_{\widetilde{S_m}}(\varepsilon)\right\|^2}{\sigma^2}$.

On a alors $\left\langle \varepsilon, \widehat{f}_{\widehat{m}} - f^* \right\rangle = \left\langle p_{\mathbb{R}f^*}(\varepsilon) + p_{\widetilde{S_{\widehat{m}}}}(\varepsilon), \widehat{f}_{\widehat{m}} - f^* \right\rangle$. En conséquence :

$$2\left\langle \varepsilon, \widehat{f}_{\widehat{m}} - f^* \right\rangle - \sigma^2 \mathrm{pen}(\widehat{m}) \leq \frac{1}{a}\left\|\widehat{f}_{\widehat{m}} - f^*\right\|^2 + a\sigma^2\left(N^2 + U_{\widehat{m}} - \frac{\mathrm{pen}(\widehat{m})}{a}\right).$$

Considérons donc désormais $Z = N^2 + U_{\widehat{m}} - \frac{\mathrm{pen}(\widehat{m})}{a}$. On a :

$$\begin{aligned}
\mathbb{E}[Z] &= 1 + \mathbb{E}\left[U_{\widehat{m}} - \tfrac{\mathrm{pen}(\widehat{m})}{a}\right] \\
&\leq 1 + \mathbb{E}\left[\sup_{m\in\mathcal{M}}\left(U_m - \tfrac{\mathrm{pen}(m)}{a}\right)^+\right] \\
&\leq 1 + \sum_{m\in\mathcal{M}} \mathbb{E}\left[\left(U_m - \tfrac{\mathrm{pen}(m)}{a}\right)^+\right].
\end{aligned}$$

Ici, l'application $F_m(\varepsilon) = \left\|p_{\widetilde{S_m}}(\varepsilon)\right\|$ est naturellement 1-lipschitzienne. En utilisant le lemme précédent, puis une inégalité de Jensen, on a :

$$\begin{aligned}
F_m(\varepsilon) &\leq \mathbb{E}\left[\left\|p_{\widetilde{S_m}}(\varepsilon)\right\|\right] + \sigma\sqrt{2\xi_m} \\
&\leq \sqrt{\mathbb{E}\left[\left\|p_{\widetilde{S_m}}(\varepsilon)\right\|^2\right]} + \sigma\sqrt{2\xi_m} \\
&= \sigma\left(\sqrt{\dim\left(\widetilde{S_m}\right)} + \sqrt{2\xi_m}\right) \\
&\leq \sigma\left(\sqrt{d_m} + \sqrt{2\xi_m}\right)
\end{aligned}$$

d'où $U_m = \frac{F_m(\varepsilon)^2}{\sigma^2} \leq \left(\sqrt{d_m} + \sqrt{2\xi_m}\right)^2$, et donc le résultat souhaité. □

Désormais, on va chercher des bons candidats de $\mathrm{pen}(m)$ pour minimiser la majoration donnée par le théorème précédent. On aimerait avoir $\rho(\mathcal{M})$ proche de 1, autrement dit $\left(R\left(d_m, \frac{\mathrm{pen}(m)}{a}\right)\right)$ proche d'une probabilité sur $\mathcal{M}$.

Corollaire 11 :

Soit π une mesure de probabilités sur $\mathcal{M}$. On définit $\mathrm{pen}(m) = K\left(\sqrt{d_m} + \sqrt{2\ln\left(\frac{1}{\pi(m)}\right)}\right)^2$. Dans ce cas, il existe C_K telle que :

$$R\left(\widehat{f}_{\widehat{m}}\right) \leq C_K \times \min_{m\in\mathcal{M}}\left[R\left(\widehat{f}_m\right) + \sigma^2\left(1 + \ln\left(\frac{1}{\pi(m)}\right)\right)\right].$$

La dernière étape consiste à bien choisir la distribution π. On veut ici obtenir une majoration $\sigma^2\ln\left(\frac{1}{\pi(m)}\right) \leq cR\left(\widehat{f}_m\right)$. Or $R\left(\widehat{f}_m\right) \geq \sigma^2 d_m$ donc il suffit de garantir $\ln\left(\frac{1}{\pi(m)}\right) \leq cd_m$, autrement dit $\pi(m) \geq \exp(-cd_m)$. Il faut donc prendre c assez grande pour que $\sum_{m\in\mathcal{M}} e^{-cd_m} \leq 1$, ce qui ne donne pas des bons résultats pour $\mathcal{M}$ trop grand.

Remarque 12 (Cas du modèle à coordonnées parcimonieuses) **:**

Ici, $\mathcal{M} = \mathcal{P}([p])$ et $S_m = \text{Vect}(X_i,\, i \in m)$.

Considérons $\pi_m = \frac{e^{-\alpha|m|}}{Z_\alpha}$. Ici, $Z_\alpha = \sum_{m\in\mathcal{M}} e^{-\alpha|m|} = \sum_{j=0}^{p} \binom{p}{j} e^{-j\alpha} = (1+e^{-\alpha})^p$. Dans ce cas $\ln\left(\frac{1}{\pi(m)}\right) = \alpha|m| + p\ln(1+e^{-\alpha})$. Pour que le terme de droite n'explose pas, il faut $\alpha \approx \ln(p)$, de sorte que $\ln\left(\frac{1}{\pi(m)}\right) \lesssim \ln(p)|m| + 1$.

Un choix alternatif est de prendre $\pi(m) = \frac{1}{Z} \times \frac{e^{-|m|}}{\binom{p}{|m|}}$ pour compenser l'explosion du nombre de candidats. Dans ce cas, $Z = \frac{e-e^{-p}}{e-1}$ est de l'ordre de 1, n'explose pas pour $p \to \infty$. Ici, $-\ln(\pi(m)) \leq |m|\left(2 + \ln\left(\frac{p}{|m|}\right)\right) + \ln\left(\frac{e-1}{e}\right)$. Dans ce cas de figure, on obtient finalement l'inégalité d'oracle :

$$R\left(\widehat{f}_{\widehat{m}}\right) \leq \widetilde{C_k} \min_{m\in\mathcal{M}} \left(R\left(\widehat{f}_m\right) + \sigma^2|m|\left(1 + \ln\left(\frac{p}{|m|}\right)\right)\right) \leq \widetilde{C_k} \times \sigma^2|m^*|\left(1+\ln\left(\frac{p}{|m|}\right)\right).$$

2.2 Minorations et critères optimaux

Soit $(\mathbb{P}_f)_{f\in\mathcal{F}}$ une famille de distributions sur $\mathbb{X}$. On suppose que l'espace de paramètres $(\mathcal{F}, d)$ est métrique. Pour tout paramètre $q > 0$, et tout estimateur $\widehat{f} : \mathbb{X} \to \mathcal{F}$ on peut définit un risque $R_f\left(\widehat{f}\right) = \mathbb{E}_{Y\sim f}\left[d\left(f, \widehat{f}(Y)\right)^q\right]$.

Pour une distribution $\mathbb{P}_f$ fixée, le meilleur estimateur est naturellement $\widehat{f} = f$, pour lequel on réalise $\min_{\widehat{f}:\mathbb{X}\to\mathcal{F}} R_f\left(\widehat{f}\right) = 0$.

Définition 13 (Risque minimax) **:**

On définit le risque minimax de la classe d'estimateurs $\mathcal{F}$ par : $R^*(\mathcal{F}) = \inf_{\widehat{f}:\mathbb{X}\to\mathcal{F}} \sup_{f\in\mathcal{F}} R_f\left(\widehat{f}\right)$.

On aimerait une minoration sur $R^*(\mathcal{F})$.

Proposition 14 (Inégalité de Fano) **:**

Soient $f_1, \ldots, f_N \in \mathcal{F}$. On notera $\mathbb{P}_j = \mathbb{P}_{f_j}$. Soit $\mathbb{Q}$ une probabilité telle que, pour tout j, on a $\mathbb{P}_j \ll \mathbb{Q}$. Par exemple, $\mathbb{Q} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{P}_j$ convient. Dans ce cas :

$$\min_{\theta:\mathbb{X}\to[N]} \max_{j\in[N]} \mathbb{P}_j(\theta \neq j) \geq 1 - \frac{1+\overline{K}}{\ln(N)},$$

où on définit $\overline{K} = \frac{1}{N}\sum_{j=1}^{n} K(\mathbb{P}_j, \mathbb{Q})$ comme la divergence de Kullback-Leibler moyenne.

Démonstration. Pour rappel, $K(\mathbb{P}_j, \mathbb{Q}) = \int_{\mathbb{X}} \ln\left(\frac{d\mathbb{P}_j(x)}{d\mathbb{Q}(x)}\right) \mathrm{d}\mathbb{P}_j(x)$.

On a naturellement $\max_{j\in[N]} \mathbb{P}_j(\theta \neq j) \geq \frac{1}{N}\sum_{j=1}^{N} \mathbb{P}_j(\theta \neq j)$, donc :

$$\min_{\theta} \max_{j} \mathbb{P}_j(\theta \neq j) \geq 1 - \max_{\theta} \frac{1}{N}\sum_{j=1}^{N} \mathbb{P}_j(\theta = j)\,.$$

On utilise désormais le lemme suivant :

Lemme 15 :
$\max_{\theta:\mathbb{X}\to[N]} \sum_{j=1}^{N} \mathbb{P}_j(\theta = j) = \mathbb{E}_{\mathbb{Q}}\left[\max_{j\in[N]} \frac{d\mathbb{P}_j}{d\mathbb{Q}}\right]$.

Démonstration. Pour cela, on utilise Fubini dans le cas positif :

$$\sum_{j=1}^{N} \mathbb{P}_j(\theta = j) = \int_{\mathbb{X}} \sum_{j=1}^{N} \mathbb{1}_{\theta(x)=j} \frac{d\mathbb{P}_j(x)}{d\mathbb{Q}(x)}\, \mathrm{d}\mathbb{Q}(x)\,.$$

Pour tout choix de θ, à $x \in \mathbb{X}$ fixé, on a au plus un des termes dans la somme qui est non nul. On en déduit une majoration par $\mathbb{E}_{\mathbb{Q}}[\ldots]$. En outre, pour $\theta(x) = \operatorname{argmax}_{j\in[N]} \frac{d\mathbb{P}_j(x)}{d\mathbb{Q}(x)}$, on obtient le cas d'égalité. □

En réinjectant ce résultat dans l'inégalité précédente, on a donc :

$$\min_{\theta} \max_{j} \mathbb{P}_j(\theta \neq j) \geq 1 - \frac{1}{N}\mathbb{E}_{\mathbb{Q}}\left[\max_{j} \frac{d\mathbb{P}_j}{d\mathbb{Q}}\right].$$

Pour toute application $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ convexe croissante, par Jensen, on a :

$$\varphi\left(\mathbb{E}\left[\max_j Z_j\right]\right) \leq \mathbb{E}\left[\varphi\left(\max_j Z_j\right)\right] = \mathbb{E}\left[\max_j \varphi(Z_j)\right] \leq \sum_j \mathbb{E}[\varphi(Z_j)]\,.$$

On va ici utiliser cette inégalité avec l'application $\varphi(x) = x\ln(x) - x + 1$ sur $[1, \infty[$ qu'on prolonge par 0 sur $[0, 1]$. On considère également $\psi(x) = x\ln(x) - x + 1$ sur $\mathbb{R}^+$, telle que $\psi \geq \varphi$.

On a donc $\varphi\left(\mathbb{E}_{\mathbb{Q}}\left[\max_j \frac{d\mathbb{P}_j}{d\mathbb{Q}}\right]\right) \leq \sum_j \mathbb{E}_{\mathbb{Q}}\left[\varphi\left(\frac{d\mathbb{P}_j}{d\mathbb{Q}}\right)\right] \leq \sum_j \mathbb{E}_{\mathbb{Q}}\left[\psi\left(\frac{d\mathbb{P}_j}{d\mathbb{Q}}\right)\right] = \sum_j K(\mathbb{P}_j, \mathbb{Q}) = N\overline{K}$.

Posons $N \times U = \mathbb{E}_{\mathbb{Q}}\left[\max_j \frac{d\mathbb{P}_j}{d\mathbb{Q}}\right]$. À j fixé, $\mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}_j}{d\mathbb{Q}}\right] = 1$ donc $NU \geq 1$ et :

$$\varphi(NU) = \psi(NU) = NU\ln(N) + N(U\ln(U) - U + 1) - (N-1) \geq NU\ln(N) - (N-1)\,,$$

d'où $NU\ln(N) \leq N\overline{K} + N - 1$, et donc $U \leq \frac{\overline{K}+1}{\ln(N)}$, le résultat voulu. □

On cherche désormais à appliquer l'inégalité de Fano pour minorer $R^*(\mathcal{F})$.

Lemme 16 :

On a la minoration :

$$\inf_{\widehat{f}:\mathbb{X}\to\mathcal{F}} \max_{j=1}^{N} \mathbb{E}_{Y\sim f_j}\left[d\Big(\widehat{f}(Y), f_j\Big)^q\right] \geq \frac{1}{2^q}\left(1 - \frac{1+\overline{K}}{\ln(N)}\right) \min_{i\neq k} d(f_i, f_k)^q \,.$$

Démonstration. Posons $\widehat{\theta}(y) = \operatorname*{argmin}_{j=1}^{N} d\Big(\widehat{f}(y), f_j\Big)$. On a alors :

$$\mathbb{1}_{\widehat{\theta}(y)\neq j} \min_{i\neq k} d(f_i, f_k) \leq d\Big(f_{\widehat{\theta}(y)}, f_j\Big) \leq d\Big(f_{\widehat{\theta}(y)}, \widehat{f}(y)\Big) + d\Big(\widehat{f}(y), f_j\Big) \leq 2d\Big(\widehat{f}(y), f_j\Big) \,.$$

Si on passe à l'espérance :

$$\max_{j=1}^{N} \mathbb{E}_{Y\sim\mathbb{P}_j}\left[d\Big(\widehat{f}(Y), f_j\Big)^q\right] \geq \frac{1}{2^q} \min_{i\neq k} d(f_i, f_k)^q \inf_{\widehat{\theta}} \max_{j=1}^{N} \mathbb{P}_j\Big(\widehat{\theta} \neq j\Big) \,.$$

et alors la minoration ne dépend plus de $\widehat{f}$, d'où le résultat en utilisant l'inégalité de Fano. □

On aimerait donc d'une part maximiser la distance entre les distributions, et d'autre part exercer un contrôle sur $\overline{K}$, par exemple $\overline{K} \leq \frac{\ln(N)}{2}$.

Par la suite, on travaille dans le cas particulier où $\mathbb{P}_f \sim \mathcal{N}(f, \sigma^2 I_n)$, $\mathcal{F} = \mathbb{R}^n$ et $d(f,g) = \|f-g\|_2$. Dans ce cas :

$$K(\mathbb{P}_f, \mathbb{P}_g) = \mathbb{E}_{Y\sim\mathbb{P}_f}\left[\ln\left(\frac{\exp\Big(-\frac{\|Y-f\|^2}{2\sigma^2}\Big)}{\exp\Big(-\frac{\|Y-g\|^2}{2\sigma^2}\Big)}\right)\right] = \mathbb{E}_f\left[\frac{\|Y-g\|^2 - \|Y-f\|^2}{2\sigma^2}\right] = \frac{\|f-g\|^2}{2\sigma^2} \,.$$

On rappelle que pour un vecteur $\beta \in \mathbb{R}^p$, on pose $\|\beta\|_0 := |\{i \in [p],\, \beta_i \neq 0\}|$.

Lemme 17 :

Pour tous $D \leq p$ fixés, il existe $\beta_1, \ldots, \beta_N \in \{0,1\}^p$ des vecteurs tels que :

1. $\forall i \in [N],\ \|\beta_i\|_0 = D$,
2. $\forall j \neq k,\ \|\beta_j - \beta_k\| > D$,
3. $\ln(N) \geq \frac{D}{2}\ln\big(\frac{p}{5D}\big)$.

Posons $f_j = r \times X\beta_j$ pour une famille β issue du lemme précédent. On définit également :

$$\underline{C_X} := \min_{\|\beta\|_0\leq 2D} \frac{\|X\beta\|_2}{\|\beta\|_2} \leq \max_{\|\beta\|_0\leq 2D} \frac{\|X\beta\|}{\|\beta\|} =: \overline{C_X} \,.$$

Dans ce cas, en particulier :

$$r^2\underline{C_X}^2 D \leq r^2\underline{C_X}^2\|\beta_j - \beta_k\|^2 \leq d(f_j, f_k)^2 \leq r^2\overline{C_X}^2\|\beta_j - \beta_k\|^2 \leq r^2\overline{C_X}^2 2D \,.$$

En outre, si $\mathbb{Q} = \mathbb{P}_1$, on a $\overline{K} = \frac{1}{N}\sum_{j=1}^{N} \frac{\|f_j - f_1\|^2}{2\sigma^2} \leq \max_{j\neq k} \frac{\|f_j - f_k\|}{2\sigma^2} \leq \frac{\overline{C_X} D r^2}{\sigma^2}$.

Pour avoir la majoration souhaitée sur $\overline{K}$, on prends $r^2 = \frac{\sigma^2 \ln(N)}{2D\overline{C_X}^2}$. On a enfin la minoration :

$$\inf_{\widehat{f}:\mathbb{X}\to\mathcal{F}} \max_{j=1}^{N} \mathbb{E}_{Y\sim f_j}\left[d\Big(\widehat{f}(Y), f_j\Big)^2\right] \geq \frac{1}{8}\left(\frac{\sigma C_X}{\overline{C_X}}\right)^2 \times \left(\frac{\ln(N)}{2} - 1\right).$$

Théorème 18 :

On a donc :

$$\inf_{\widehat{f}} \sup_{\|\beta\|_0 = D} \mathbb{E}_{f=X\beta}\left[\left\|\widehat{f} - X\beta\right\|^2\right] \geq \frac{1}{8}\left(\frac{\sigma C_X}{\overline{C_X}}\right)^2 \left(\frac{D}{4}\ln\Big(\frac{p}{5D}\Big) - 1\right).$$

Démonstration. La partie de gauche vient du fait qu'on prend un sup sur tous les β qui ont D coefficients non-nuls, plus grand que le max sur les N vecteurs $r\beta_i$ en particulier. La partie de droite découle du dernier point du lemme donnant la famille β, à savoir la minoration de $\ln(N)$. □

Si on compare cette minoration à la majoration sur l'estimateur précédent, on constate que les bornes sont du même ordre de grandeur, donc que notre estimateur dans le cas parcimonieux est optimal à une constante multiplicative près.

3 Convexification

Le souci de l'approche précédente est que les calculs ont souvent une complexité prohibitive.

Ainsi, si $\mathcal{M} = \mathcal{P}([p])$, alors a priori on a au moins $|\mathcal{M}| = 2^p$ opérations à effectuer.

Tant que $\mathcal{M}$ reste relativement petit, que les colonnes (X_i) sont orthogonales, ou qu'on considère un estimateur constant par morceaux, on peut trouver des méthodes pour rentabiliser les calculs et atteindre une complexité polynomiale.

Une façon de contourner le problème est d'utiliser des algorithmes d'approximation gloutons, qui se déplacent dans $\mathcal{M}$ jusqu'à atteindre un « maximum local ». On va ici considérer une autre approche : la convexification.

3.1 Cas parcimonieux

On considère $\mathcal{M} = \mathcal{P}([p])$, avec $\pi_m = e^{-|m|\ln(p)}$. En considérant la pénalité induite par π_m, on se ramène donc à $\widehat{m} = \operatorname*{argmin}_{m\in\mathcal{M}} \left\|Y - \widehat{f}_m\right\|^2 + \lambda|m|$, où $\lambda = K\sigma^2\Big(1 + \sqrt{2\ln(p)}\Big)^2$. Si on pose $\widehat{\beta}_m = \operatorname*{argmin}_{\mathrm{Supp}(\beta)=m} \|Y - X\beta_m\|^2$, alors le vecteur aléatoire $\widehat{\beta}_{\widehat{m}}$ minimise le risque empirique.

Pour minimiser $\|Y - X\beta\|^2 + \lambda\|\beta\|_0$ sur $\mathbb{R}^p$, un souci se pose car $\beta \mapsto \|\beta\|_0$ n'est pas convexe. L'enveloppe convexe de $\{\beta,\ \|\beta\|_0 \leq D\}$ est $\mathbb{R}^p$. Si on se restreint à une boule $B_{\|.\|_2}(0,R)$, l'enveloppe convexe devient $B_{\|.\|_1}(0,R)$.

Définition 19 (Estimateur Lasso) **:**
Il est alors naturel de considérer $\widehat{\beta}_\lambda = \text{argmin}_{\beta\in\mathbb{R}^p} \mathcal{L}_\lambda(\beta)$, où $\mathcal{L}_\lambda(\beta) = \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$.

Cette application est convexe, mais n'est pas lisse, ce qui jouera un rôle crucial.

Définition 20 (Problème dual) **:**
En considérant $R = \left\|\widehat{\beta}_\lambda\right\|_1$, on a $\widehat{\beta_\lambda} = \underset{\|\beta\|_1\leq R}{\text{argmin}}\|Y - X\beta\|^2$.

On va donc s'intéresser à la minimisation de cette norme sur un convexe.

Graphiquement, on constate que si R est trop grand, alors on va simplement être au centre de l'ellipse, puis se déplacer le long d'une droite au fur et à mesure que R diminue, jusqu'à atteindre un hyperplan d'axes. Si R est assez petit, alors le minimum est alors atteint dans un coin du domaine, donc de fait on élimine certaines coordonnées de β. On opère ainsi de la sélection de variables.

Définition 21 (Ensemble sous-différentiel) **:**
Pour une fonction convexe $F : \mathbb{R}^p \to \mathbb{R}$ non nécessairement lisse, on pose :

$$\partial F(x) = \{\omega \in \mathbb{R}^p,\ \forall y \in \mathbb{R}^p,\ F(y) - F(x) \geq \langle \omega, y - x\rangle\} \neq \emptyset\,.$$

On dit que $\omega \in \partial F(x)$ est un sous-gradient de F. Si F est différentiable en x, $\partial F(x) = \{\nabla F_x\}$.

Lemme 22 :
Si F est convexe, on a les propriétés suivantes :

1. Monotonie : $\forall x \in \partial F(x),\ \forall y \in \partial F(y),\ \langle \omega_y - \omega_x, y - x\rangle \geq 0$,
2. Optimalité : x minimise F ssi $0 \in \partial F(x)$.

Démonstration. Pour le premier point, on a $F(y) - F(x) \geq \langle \omega_x, y - x\rangle$ par définition de $\partial F(x)$, et de même $F(x) - F(y) \geq \langle -\omega_y, y - x\rangle$, d'où le résultat en prenant la somme. □

Remarque 23 (Cas de $\|.\|_1$) **:**
Si $F(x) = \|x\|_1$, alors $\partial F(x) = \{z \in \mathbb{R}^p,\ \|z\|_\infty \leq 1,\ \langle z, x\rangle = \|x\|_1\}$. Autrement dit, on a $z_j = \text{sign}(x_j) \in \{\pm 1\}$ lorsque $x_j \neq 0$, et $z_j \in [-1, 1]$ sinon.

Par optimalité, comme $\widehat{\beta}_\lambda$ minimise $\mathcal{L}_\lambda$, on a $\widehat{z} \in \partial\left\|\widehat{\beta}_\lambda\right\|_1$ tel que $X^TX\widehat{\beta}_\lambda = X^TY - \frac{\lambda}{2}\widehat{z}$.

Lorsque les colonnes de X sont orthogonales, X^TX est l'identité, alors pour tout j on a $\widehat{\beta}_{\lambda,j} = 0$ ou bien $\widehat{\beta}_{\lambda,j} = X^TY - \frac{\lambda}{2}\text{sign}\left(\widehat{\beta}_{\lambda,j}\right)$.

En particulier, pour avoir $\widehat{\beta}_{\lambda,j}=0$, il faut nécessairement avoir $\left|X_j^TY\right|\leq\frac{\lambda}{2}$. Dans ce cas, on peut plus largement prendre $\widehat{\beta}_{\lambda,j}=X_j^TY\times\left(1-\frac{\lambda/2}{|X_j^TY|}\right)^+$.

3.2 Performances de l'estimateur Lasso

Théorème 24 :

On rappelle que $Y=X\beta^*+\varepsilon$. Si $\lambda>3\left\|X^T\varepsilon\right\|_\infty$, alors :

$$\left\|X\widehat{\beta}-X\beta^*\right\|_2^2\leq\inf_{\beta\neq0}\left(\|X\beta-X\beta^*\|_2^2+\frac{\lambda^2}{K(\beta)^2}\|\beta\|_0\right).$$

En notant $S=\mathrm{Supp}(\beta)$, on définit la constante de compatibilité par :

$$K(\beta)=\inf\left\{\sqrt{\|\beta\|_0}\frac{\|Xu\|_2}{\|u|_S\|_1},\ u\in\mathcal{C}(\beta)\right\},$$

où $\mathcal{C}(\beta)=\{u,\ \|u|_{S^c}\|_1<5\|u|_S\|_1\}$.

Démonstration. Par définition de $\widehat{\beta}$, $0\in\partial\mathcal{L}_\lambda\left(\widehat{\beta}\right)$ d'où $\widehat{z}\in\partial\left\|\widehat{\beta}\right\|_1$ tel quel $2X^T\left(Y-X\widehat{\beta}\right)=\lambda\widehat{z}$.

Pour $\beta\in\mathbb{R}^p$ fixé, en passant au produit scalaire avec $\widehat{\beta}-\beta$, on a $\left\langle\widehat{z},\widehat{\beta}-\beta\right\rangle\geq\left\langle z,\widehat{\beta}-\beta\right\rangle$ pour tout $z\in\partial\left\|\widehat{\beta}\right\|_1$. On a donc :

$$2\left\langle X\left(\widehat{\beta}-\beta^*\right),X\left(\widehat{\beta}-\beta\right)\right\rangle\leq2\left\langle X^T\varepsilon,\widehat{\beta}-\beta\right\rangle-\lambda\left\langle z,\widehat{\beta}-\beta\right\rangle.$$

On pose par la suite $\mathcal{A}(\beta)=2\left\langle X\left(\widehat{\beta}-\beta^*\right),X\left(\widehat{\beta}-\beta\right)\right\rangle$.

On utilise alors le lemme suivant. Plus précisément, en utilisant la formule d'Al-Kashi, on a $\mathcal{A}(\beta)=\left\|X\left(\widehat{\beta}-\beta^*\right)\right\|^2+\left\|X\left(\widehat{\beta}-\beta\right)\right\|^2-\|X(\beta-\beta^*)\|^2$. En conséquence, si $\mathcal{A}(\beta)\leq0$, alors en particulier $\left\|X\left(\widehat{\beta}-\beta^*\right)\right\|^2\leq\|X(\beta-\beta^*)\|^2$. Sinon, si $\mathcal{A}(\beta)>0$, on a :

$$\begin{aligned}\left\|X\left(\widehat{\beta}-\beta^*\right)\right\|_2^2+\left\|X\left(\widehat{\beta}-\beta\right)\right\|_2^2&\leq\|X(\beta-\beta^*)\|_2^2+2\lambda\left\|\left(\widehat{\beta}-\beta\right)|_S\right\|_1\\&\leq\|X(\beta-\beta^*)\|_2^2+2\lambda\frac{\sqrt{\|\beta\|_0}}{K(\beta)}\times\left\|X\left(\widehat{\beta}-\beta\right)\right\|_2\\&\leq\|X(\beta-\beta^*)\|_2^2+\lambda^2\frac{\|\beta\|_0}{K(\beta)^2}+\left\|X\left(\widehat{\beta}-\beta\right)\right\|_2^2\end{aligned}$$

car $2ab\leq a^2+b^2$, d'où enfin la majoration souhaitée pour tout $\beta\neq0$. □

Lemme 25 :

On a les résultats suivants :

— $\mathcal{A}(\beta)\leq2\lambda\left\|\left(\widehat{\beta}-\beta\right)|_S\right\|_1$.

— Si $\mathcal{A}(\beta) > 0$, alors $\widehat{\beta} - \beta \in \mathcal{C}(\beta)$.

Démonstration. Pour la majoration, on a :

$$\begin{aligned}\mathcal{A}(\beta) &\leq 2\left\|X^T\varepsilon\right\|_\infty \left\|\widehat{\beta} - \beta\right\|_1 - \lambda\left\langle z|_S, \left(\widehat{\beta} - \beta\right)|_S\right\rangle - \lambda\left\langle z|_{S^c}, \left(\widehat{\beta} - \beta\right)|_{S^c}\right\rangle \\ &\leq \tfrac{2}{3}\lambda\left\|\widehat{\beta} - \beta\right\|_1 + \lambda\left(\left\|\left(\widehat{\beta} - \beta\right)|_S\right\|_1 - \left\|\left(\widehat{\beta} - \beta\right)|_{S^c}\right\|_1\right),\end{aligned}$$

d'où l'inégalité souhaitée en oubliant les termes restreints à S^c.

Pour le second point, si $\mathcal{A}(\beta) > 0$, alors dans l'inégalité ci-dessus, si on met les termes restreints à S^c à gauche de l'inégalité, on obtient bien $\widehat{\beta} - \beta \in \mathcal{C}(\beta)$. □

Corollaire 26 :

Supposons que les p colonnes de X sont normées, que $\|X_j\| = 1$. Si $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, pour $\lambda = 3\sigma\sqrt{2K\ln(p)}$, alors avec probabilité au moins $1 - \frac{1}{p^{K-1}}$, on a :

$$\begin{aligned}\left\|X\widehat{\beta} - X\beta^*\right\|_2^2 &\leq \inf_\beta\left(\|X\beta - X\beta^*\|_2^2 + \tfrac{18K\sigma^2\|\beta\|_0\ln(p)}{K(\beta)^2}\right) \\ &\leq \inf_m\left(\left\|X\widehat{\beta}_m - X\beta^*\right\|_2^2 + \tfrac{18K\sigma^2|m|\ln(p)}{K(\beta)^2}\right)\end{aligned}$$

Démonstration. Il suffit de prouver que $\lambda > 3\left\|X^T\varepsilon\right\|$ avec la probabilité souhaitée. En effet, une fois ceci fait, le reste n'est que substitution de termes, puis passage à un sous-ensembles. On peut majorer par une somme de queues de gaussiennes centrées réduites :

$$\mathbb{P}\left(\lambda \leq 3\left\|X^T\varepsilon\right\|_\infty\right) \leq \sum_{j=1}^p \mathbb{P}\left(\frac{|\langle X_j, \varepsilon\rangle|}{\sigma} \geq \sqrt{2K\ln(p)}\right) \leq \sum_{j=1}^p \exp\left(-\frac{1}{2}2K\ln(p)\right) = \frac{1}{p^{K-1}}$$

d'où le résultat souhaité. □

3.3 Biais de l'estimateur

Le souci de l'estimateur est qu'il a tendance à diminuer les coefficients. Une approche pour contourner ce souci, appelée *Gauss-Lasso*, consiste à utiliser Lasso pour obtenir $\widehat{\beta}_\lambda$, prendre son support $\widehat{m}_\lambda = \mathrm{Supp}\left(\widehat{\beta}_\lambda\right)$, puis l'estimateur des moindres carrés $\widehat{f}_{\widehat{m}_\lambda}$ sur le modèle sélectionné.

Une autre approche, appelée *adaptive-Lasso*, part du constat qu'on a remplacé $\|\beta\|_0$ par $\|\beta\|_1$. Or, pour β au voisinage de $\widehat{\beta}^{init}$, on a $\sum_{j\in\mathrm{Supp}\left(\widehat{\beta}^{init}\right)}\left|\frac{\beta_j}{\widehat{\beta}_j^{init}}\right| \approx \|\beta\|_0$. En partant d'un estimateur initial correct, on peut donc itérer :

$$\widehat{\beta}^{i+1} = \underset{\beta}{\mathrm{argmin}}\left(\|Y - X\beta\|^2 + \lambda\sum_{j\in\mathrm{Supp}\left(\widehat{\beta}^i\right)}\left|\frac{\beta_j}{\widehat{\beta}_j^i}\right|\right),$$

pour raffiner notre estimateur.

3.4 Calcul effectif de l'estimateur

On peut par exemple faire une descente de gradient à pas optimal, en faisant un cycle coordonnée par coordonnée. Lorsque $\|X_j\|_2 = 1$ et $\beta_j \neq 0$, en posant $R_j = \left\langle X_j, Y - \sum_{k \neq j} \beta_k X_k \right\rangle$ on a :

$$\partial_j \mathcal{L}\lambda(\beta) = -2(R_j - \beta_j) + \lambda\sigma(\beta_j)$$

donc $\beta_j^* = R_j\left(1 - \frac{\lambda}{2|R_j|}\right)^+$. Cette méthode donne rapidement des bonnes approximations, mais converge lentement vers la valeur exacte.

Il existe une autre approche, plus algébrique. On remarque ici qu'on peut restreindre l'égalité au support $\widehat{m}_\lambda$ de $\widehat{\beta}_\lambda$, de sorte que :

$$X_{\widehat{m}}^T X_{\widehat{m}}^T \widehat{\beta}_\lambda = X_{\widehat{m}}^T Y - \frac{\lambda}{2}\mathrm{Sign}\left(\widehat{\beta}_\lambda\right)$$

et, comme on impose le support $\widehat{m}$, le signe est bien défini dans $\{\pm 1\}$. Dans ce cas, on a donc $\widehat{\beta}_\lambda = \left(X_{\widehat{m}}^T X_{\widehat{m}}\right)^{-1}\left(X_{\widehat{m}}^T Y - \frac{\lambda}{2}\mathrm{Sign}\left(\widehat{\beta}_\lambda\right)\right)$. L'application $\lambda \mapsto \widehat{\beta}_\lambda$ est donc continue, affine par morceaux. Avec plus de travail, on peut alors expliciter les points de jonction (λ_i), et les estimateurs associés.

3.5 Autres méthodes de parcimonie

Une approche générale consiste à adjoindre à $\|Y - X\beta\|_2^2$ une pénalité $\lambda\Omega(\beta)$, où les singularités de Ω dans $\{\beta,\ \Omega(\beta) \leq 1\}$ correspondent aux *patterns* qu'on cherche à sélectionner.

Par exemple, on a vu que dans le cas de coordonnées parcimonieuses, on transforme l'objet qui nous intéresse, $\sum \mathbb{1}_{\beta_j \neq 0}$, en $\sum|\beta_j|$, qui a ses singularités alignées sur la première fonction. Plus largement, dans le cas de la parcimonie par blocs, on transforme $\sum_{k=1}^{r} \mathbb{1}_{\beta|_{G_k} \neq 0}$ en $\sum_{k=1}^{r} \|\beta|_{G_k}\|_2$.

4 Modèles graphiques

On cherche ici des bons moyens de représenter l'indépendance conditionnelle entre des variables aléatoires, pour mieux comprendre les relations de causalité entre variables.

Simplement observer les corrélations entre variables ne suffit pas. Pour savoir de qui dépend Y parmi (X_i), on peut faire une régression linéaire, une sélection de variables.

Cependant, il reste un souci. Si on montre que $Y \cong f(X_1)$, X_1 n'est pas forcément indépendant des autres variables. Si $X_1 \cong (X_2)$, alors changer X_2 impactera fortement Y, ce qui n'est pas donné par la seule fonction f.

4.1 Modélisation par équations structurelles (SEM)

Pour aller un cran plus loin, on considère p variables (X_i), et un graphe acyclique $\vec{g}$ orienté, qui induit une structure causale.

On note $pa(i) = \{j,\ j \to i\}$ les parents du sommet i et $de(i) = \{j, i \to^* j\}$ les descendants de i (dont i). Enfin, si $S \subset [p]$, on note $x_S = (x_i)_{i \in S}$.

Définition 27 (SEM) **:**
On suppose qu'il existe, pour chaque $a \in [p]$, un bruit ε_a et une fonction F_a telles que $X_a = F_a\big(X_{pa(a)}, \varepsilon_a\big)$.

Remarque 28 (Interventions, Judea Pearl) **:**
On veut pouvoir quantifier l'effet de la variable X_i sur X_j, sans considérer la corrélation de X_i et des variables $X_{pa(i)}$ dont elle est issue.

Pour ce faire, on coupe toutes les flèches de $pa(i)$ à i dans $\vec{g}$, et on remplace F_i par une constante x_i.

Remarque 29 :
Conditionnellement à $X_{pa(i)}$, X_i est indépendant de $(X_j)_{j \notin de(i)}$.

4.2 Rappels sur l'indépendance conditionnelle

Définition 30 :
On dit que X et Y sont indépendantes conditionnellement à Z si la mesure (aléatoire)

$\mathcal{L}(X, Y|Z)$ est égale à $\mathcal{L}(X|Z) \otimes \mathcal{L}(Y|Z)$.

Lemme 31 :

Supposons que le triplet (X, Y, Z) est à densité $f > 0$ pour une mesure produit σ-finie (Lebesgue par exemple).

Alors X et Y sont indépendantes conditionnellement à Z ssi $f(x, y|z) = f(x|z)f(y|z)$ ssi $f(x, y, z) = f(x|z)f(y, z)$ ssi $f(x|y, z) = f(x|z)$.

Lemme 32 :

Si X est indépendant de (Y, W) conditionnellement à Z, alors X est indépendant de Y conditionnellement à (W, Z).

4.3 Modèles graphiques orientés

Définition 33 :

Les variables $(X_1, \dots, X_p)$ suivent un modèle graphique sur $\vec{g}$ lorsque, pour tout $a \in [p]$, X_a est indépendant des $(X_b)_{b \notin de(a)}$ conditionnellement à $X_{pa(a)}$. On note alors $\mathcal{L}(X) \sim \vec{g}$.

Lemme 34 :

Si $\vec{g_1} \subset \vec{g_2}$ et $\mathcal{L}(X) \sim \vec{g_1}$, alors $\mathcal{L}(X) \sim \vec{g_2}$.

Corollaire 35 :

En général, on n'a pas unicité du graphe $\vec{g}$ adapté à X.

Remarque 36 :

Il n'y a même pas unicité du graphe minimal pour l'inclusion.

Par exemple, si $X_{i+1} = X_i + \varepsilon_i$ où les ε sont des bruits iid, et $X_0 = 0$, on peut vérifier que les graphes $0 \to \cdots \to p$ et $0 \leftarrow \cdots \leftarrow p$ conviennent.

Cet exemple montre en particulier que le sens des flèches ne traduit pas nécessairement une relation causale entre variables.

Proposition 37 (Formule de factorisation) **:**

Si X à densité $f > 0$ et $\mathcal{L}(X) \sim \vec{g}$, alors $f = \prod_{a \in [p]} f\big(x_a|x_{pa(a)}\big)$.

Démonstration. Quitte à permuter, p est une feuille de $\vec{g}$. On se retrouve avec la factorisation $f = f(x_p|x_1, \dots, x_{p-1})f(x_1, \dots, x_{p-1}) = f\big(x_p\big|x_{pa(p)}\big)f(x_1, \dots, x_{p-1})$.

On conclut alors par induction sur des graphes de taille décroissante, en remontant jusqu'aux racines. □

4.4 Modèles graphiques non orientés

On considère désormais g non orienté, et $ne(a) = \{b,\ a \leftrightarrow b\}$ les voisins du sommet a, et alors $cl(a) = ne(a) \cup \{a\}$.

Définition 38 :
On dit que X suit un modèle graphique sur g lorsque, pour tout $a \in [p]$, X_a est indépendant de $(X_b)_{b \notin cl(a)}$ conditionnellement à $X_{ne(a)}$. On note alors $\mathcal{L}(X) \sim g$.

Lemme 39 :
Si $g \subset g'$ et $\mathcal{L}(X) \sim g$ alors $\mathcal{L}(X) \sim g'$.

Proposition 40 :
Si X à densité $f > 0$, alors il existe un unique graphe g^*, minimal pour l'inclusion, tel que $\mathcal{L}(X) \sim g^*$.

Définition 41 (Moralisation) **:**
On peut *moraliser* un graphe orienté $\vec{g}$ en un graphe non orienté $g^m = moral(\vec{g})$.

Pour ce faire, on désoriente les arêtes de $\vec{g}$ dans g^m, et on fait cliquer chaque ensemble $pa(a)$.

Proposition 42 :
Si X à densité $f > 0$, et $\mathcal{L}(X) \sim \vec{g}$, alors $\mathcal{L}(X) \sim g^m$.

Démonstration. Comme X est à densité $f > 0$ on peut utiliser la formule de factorisation. Alors pour tout $a \in [p]$ on a :

$$f = f\left(x_a | x_{pa(a)}\right) \times \prod_{b,\, a \in pa(b)} f\left(x_b | x_{pa(b)}\right) \times \prod_{b \neq a,\, a \notin pa(b)} f\left(x_b | x_{pa(b)}\right).$$

La partie de gauche ne dépend que des parents de x_a, de ses enfants et de leurs parents, autrement dit de $cl(a)$ (sa classe dans g^m). Le terme de droite ne fait jamais intervenir a.

En conséquence, on peut factoriser $f = g\left(x_a, x_{ne(a)}\right) \times h(x_{-a})$, où x_{-a} correspond à x privé de sa coordonnée a. Conditionnellement à $x_{ne(a)}$, on a donc bien factorisé la densité en un terme fonction de x_a et un terme fonction des autres variables non fixées, d'où l'indépendance conditionnelle, $\mathcal{L}(X) \sim g^m$. □

Proposition 43 (Formule de factorisation de Hammusley-Clifford) **:**
Soit X à densité $f > 0$. On a $\mathcal{L}(X) \sim g$ ssi $f = \prod_{c \in \text{clique}(g)} \Phi_c(x_c)$.

Démonstration. Preuve par inversion de Möbius, voir *Graphical models*, Lauritzen. □

En pratique, on considère des cas où le graphe g est simple (un arbre par exemple), ou bien on fait des hypothèses sur la loi de X.

4.5 Modèles graphiques gaussiens (GGM)

On suppose ici $X \sim \mathcal{N}(0, \Sigma)$.

4.5.1 Préliminaires

Lemme 44 (Conditionnement gaussien) **:**
Quitte à permuter les coordonnées, $A = [k]$ et $B = [\![k+1, p]\!]$. On suppose que Σ est inversible, et on note :

$$K = \Sigma^{-1} = \begin{pmatrix} K_{A,A} & K_{A,B} \\ K_{B,A} & K_{B,B} \end{pmatrix}$$

la matrice par blocs obtenue. Alors $\mathcal{L}(X_A | X_B) \sim \mathcal{N}\left(-K_{A,A}^{-1} K_{A,B} X_B, (K_{A,A})^{-1}\right)$. Autrement dit, il existe $\varepsilon_A \sim \mathcal{N}\left(0, K_{A,A}^{-1}\right)$ indépendant de X_B, tel que $X_A = K_{A,A}^{-1} K_{A,B} X_B + \varepsilon_A$.

Remarque 45 :

On a :

$$\mathrm{Cor}(X_a, X_b | X_c, c \notin \{a,b\}) := \frac{\mathrm{Cov}(X_a, X_b | X_c, c \notin \{a,b\})}{\sqrt{\mathrm{Var}(X_a|X_c, c \notin \{a,b\})\mathrm{Var}(X_b|X_c, c \notin \{a,b\})}} = -\frac{K_{a,b}}{\sqrt{K_{a,a}K_{b,b}}}$$

Démonstration. On utilise le lemme précédent, avec $A = \{a,b\}$ et B son complémentaire. La valeur de X_B n'influe que sur la moyenne, pas sur la variance de X_A. Il suffit donc de calculer $K^{-1}_{A,A}$.

Comme $K^{-1}_{A,A} = \frac{1}{\det(K_{A,A})} \begin{pmatrix} K_{b,b} & -K_{a,b} \\ -K_{a,b} & K_{a,a} \end{pmatrix}$, à une constante multiplicative près, la covariance est proportionnelle à $-K_{a,b}$ et les variances à $K_{b,b}$ et $K_{a,a}$, d'où le résultat. □

Proposition 46 :

On définit le graphe non orienté g_K, muni des arêtes $\{a,b\}$ lorsque $K_{a,b} \neq 0$.

Alors $\mathcal{L}(X) \sim g_K$, et g_K est le graphe minimal pour l'inclusion.

En outre, $X_a = \varepsilon_a - \sum_{b \in ne(a)} \frac{K_{a,b}}{K_{a,a}} X_b$, avec $\varepsilon_a \sim \mathcal{N}\left(0, \frac{1}{K_{a,a}}\right)$ indépendamment de $X_{ne(a)}$.

Démonstration. Soient $B = ne(a)$ et $A = \{a\} \sqcup cl(a)^c$ son complémentaire. Montrons que X_a et $X_{cl(a)^c}$ sont indépendants conditionnellement à X_B. Cela découle de la définition de g_K, car $c \in cl(a)^c$ ssi $c \notin ne(a)$ (et $c \neq a$) ssi $K_{a,c} = 0$. En conséquence, $K_{A,A}$ est diagonale par blocs, avec un bloc $K_{a,a}$ de dimension 1 et un bloc associé à $cl(a)^c$. En passant à l'inverse, On a donc bien l'indépendance conditionnelle souhaitée, donc $\mathcal{L}(X) \sim g_K$.

Réciproquement, supposons $\mathcal{L}(X) \sim g$. Avec A et B définis comme précédemment, on a $K^{-1}_{A,A}$ diagonale par blocs donc $K_{A,A}$ aussi. Si $\{a,b\}$ n'est pas arête dans g, alors $K_{a,b} = 0$. Autrement dit, on a bien $g_K \subset g$.

Reste à exhiber une formule pour X_a. Pour cela, on considère $A = \{a\}$ cette fois-ci. Alors $\mathcal{L}(X_a|X_B) \sim \mathcal{N}\left(-\frac{K_{a,B}}{K_{a,a}} X_B, \frac{1}{K_{a,a}}\right)$ et quitte à oublier les termes nuls, $K_{a,B}X_B = \sum_{b \in ne(a)} K_{a,b}X_b$. □

4.5.2 Estimation

On ne connait pas Σ, mais on a un échantillon $X^{(i)} \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma)$ de taille n. On souhaite alors estimer g_K.

L'approche naïve consiste à estimer K puis prendre le graphe induit. La vraisemblance de K est :

$$L(K) = \prod_{i=1}^{n} \sqrt{\frac{\det(K)}{(2\pi)^p}} e^{-\frac{1}{2}\left(X^{(i)}\right)^T K X^{(i)}}.$$

et donc $-\ln(L(K)) = \frac{np}{2}\ln(2\pi) + \frac{1}{2}\sum_{i=1}^{n}\left\langle X^{(i)}, KX^{(i)}\right\rangle_F - \frac{n}{2}\det(K)$, où $\langle A, B\rangle_F$ est associé à la norme de Frobenius, la norme euclidienne sur $\mathbb{R}^{p^2}$. Si on note $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\left(X^{(i)}\right)^T X^{(i)}$, on s'est ramené à $-\ln(L(K)) = \frac{n}{2}\Big(\Big\langle\widehat{\Sigma}, K\Big\rangle_F - \ln(\det(K))\Big) + \frac{np}{2}\ln(2\pi)$. Lorsque $\widehat{\Sigma}$ est inversible, ce qui nécessite $n \geq p$, la matrice qui maximise la vraisemblance est alors $\widehat{K} = \widehat{\Sigma}^{-1}$. Dans ce cas, la matrice $\widehat{K}$ est presque-sûrement pleine. On peut cependant en déduire un graphe $\widehat{g}$ en mettant une arête $\{a, b\}$ lorsque $\left|\widehat{K}_{a,b}\right| \geq \tau_{a,b}$ dépasse un certain seuil. Cette approche a pour autre défaut d'être assez instable numériquement, même si $n \gg p$.

Assez naturellement, puisqu'on veut obtenir un graphe relativement régulier, on pourrait plutôt opérer une sélection de variables.

Définition 47 (Estimateur graphical-lasso) **:**

On peut ainsi considérer l'estimateur :

$$\widehat{K} \in \underset{K\in S_n^+}{\operatorname{argmin}}\left(\frac{n}{2}\Big(\Big\langle\widehat{\Sigma}, K\Big\rangle - \ln(\det(K))\Big) + \lambda\sum_{a\neq b}|K_{a,b}|\right).$$

Lemme 48 :

L'application $\Phi : K \mapsto -\ln(\det(K))$ est convexe sur S_n^+.

Démonstration. Il suffit de montrer le résultat pour $K, S \in S_n^{++}$ et $\lambda \in]0, 1[$, puis de conclure par densité. Dans ce cas :

$$\begin{aligned}\Phi(\lambda K + (1-\lambda)S) &= \Phi\Big(\sqrt{K}\Big(\lambda I + (1-\lambda)\sqrt{K}^{-1}S\sqrt{K}^{-1}\Big)\sqrt{K}\Big)\\ &= \Phi(K) + \Phi\Big(\lambda I + (1-\lambda)\sqrt{K}^{-1}S\sqrt{K}^{-1}\Big).\end{aligned}$$

Quitte à diagonaliser $\sqrt{K}^{-1}S\sqrt{K}^{-1}$ dans une base orthogonale, avec des valeurs propres (σ_k), On a :

$$\begin{aligned}\Phi(\lambda K + (1-\lambda)S) &= \Phi(K) + \sum_{k=1}^{n} -\ln(\lambda + (1-\lambda)\sigma_k)\\ &\leq \Phi(K) - (1-\lambda)\sum_{k=1}^{n}\ln(\sigma_k)\\ &= \Phi(K) + (1-\lambda)\Phi\Big(\sqrt{K}^{-1}S\sqrt{K}^{-1}\Big)\\ &= \Phi(K) + (1-\lambda)(\Phi(S) - \Phi(K))\end{aligned}$$

d'où le résultat souhaité. □

On cherche donc à minimiser une fonctionnelle convexe sur un domaine convexe. Cependant, les calculs de gradients ne sont pas faciles, et requièrent des valeurs de p au plus de l'ordre de quelques milliers. L'estimateur, bien que similaire à Lasso, est difficile à interpréter en théorie, et ses performances sont correctes mais sans plus en pratiques.

Une autre approche est d'utiliser la forme $X_a = \sum_{b\neq a}\theta^*_{b,a}X_b + \varepsilon_a$, avec $\theta^*_{a,b} = -\frac{K_{a,b}}{K_{a,a}}$. En conséquence,

$\theta^* \in \underset{\theta\in M_p(\mathbb{R}),\,\mathrm{diag}(\theta)=0}{\mathrm{argmin}} \mathbb{E}\Big[\big\|X-\theta^T X\big\|_2^2\Big]$. En outre, $\big\|X-\theta^T X\big\|_2 = \big\|X^T(I-\theta)\big\|_2$. On considère alors :

$$\widehat{\theta} \in \underset{\theta\in M_p(\mathbb{R}),\,\mathrm{diag}(\theta)=0}{\mathrm{argmin}} \frac{1}{n}\sum_{i=1}^{n}\Big\|\big(X^{(i)}\big)^T(I-\theta)\Big\|_2^2 + \lambda\Omega(\theta)\,,$$

où Ω est une fonction de pénalité. Si on note $\mathbb{X} = \begin{pmatrix}\big(X^{(1)}\big)^T\\ \vdots \\ \big(X^{(n)}\big)^T\end{pmatrix}$, alors on a l'écriture alternative $\widehat{\theta} \in \underset{\theta\in M_p(\mathbb{R}),\,\mathrm{diag}(\theta)=0}{\mathrm{argmin}} \|\mathbb{X}(I-\theta)\|_F^2 + \lambda\Omega(\theta)$ avec $\|.\|_F$ la norme de Frobenius.

Si on prend la pénalité $\Omega_1(\theta) = \|\theta\|_1$, quitte à décomposer $\mathbb{X} = (\mathbb{X}_a)$ en colonnes, et de même pour $\theta = (\theta_a)$, alors $\widehat{\theta}_{\Omega_1} \in \underset{\theta\in M_p(\mathbb{R}),\,\mathrm{diag}(\theta)=0}{\mathrm{argmin}} \sum_{a=1}^{n}\big(\frac{1}{n}\|\mathbb{X}_a - \mathbb{X}\theta_a\|_2^2 + \lambda\|\theta_a\|_1\big)$. On peut calculer séparément chaque vecteur $\widehat{\theta}_a$ par Lasso. L'implémentation est relativement facile, mais on perd alors la symétrie des 0, et il n'y a pas de façon naturelle de décider si $\{a,b\}$ est une arête lorsque $\theta_{a,b} = 0 \neq \theta_{b,a}$.

La réponse naturelle à cette remarque est d'utiliser un estimateur de Lasso par groupes, en appairant $\theta_{a,b}$ et $\theta_{b,a}$, avec $\Omega_2(\theta) = \sum_{a<b}\sqrt{\theta_{a,b}^2+\theta_{b,a}^2}$. On impose ainsi la symétrie des zéros de θ, mais au prix d'un système non séparable, on ne peut plus calculer chaque colonne de $\widehat{\theta}$ à part, ce qui rajoute un facteur p dans la complexité globale.

4.6 Au-delà du cas gaussien

Dans le cas général, l'étude est très difficile. Il existe des méthodes lorsque g^* est un arbre. On peut tout de même étendre les résultats précédents aux copules gaussiennes.

Supposons que X est à densité f_X. Alors $F_a(u) = \mathbb{P}(X_a \leq u)$ est dérivable, à dérivée positive. Alors $F_a(X_a) \sim \mathcal{U}([0,1])$. Si on note Φ la fonction de répartition de $\mathcal{N}(0,1)$, on a alors la variable $Z_a = \Phi^{-1}\circ F_a(X_a) \sim \mathcal{N}(0,1)$.

On dit que X est une copule gaussienne lorsque $Z \sim \mathcal{N}(0,\Lambda)$ est un vecteur gaussien. On peut alors étendre les résultats gaussiens sur Z ci-dessus à X.

Lemme 49 :
Les variables X et Z ont le même graphe optimal.

Démonstration. Notons $h_a = \Phi^{-1}\circ F_a$. Comme $z_a = h_a(x_a)$, on a le changement de variables $f_Z(z)\,\mathrm{d}z = f_Z(h_1(x_1),\dots,h_p(x_p))\prod_{i=1}^{p} h_i'(x_i)\,\mathrm{d}x$. En conséquence, f_Z et f_X admettent les mêmes factorisations, donc le même graphe minimal. □

Le souci est ici qu'on ne connait *pas* la distribution f_X, donc on ne connait pas F_a. On pourrait commencer par estimer F_a avant de calculer Z_a. Alternativement, on utilise astucieusement le fait que $\Lambda_{a,b} = \sin\big(\frac{\pi}{2}\tau_{a,b}\big)$, où $\tau_{a,b} = \mathbb{E}_{Z,\widetilde{Z}}\Big[\mathrm{sign}\Big(\Big(Z_a-\widetilde{Z}_a\Big)\Big(Z_b-\widetilde{Z}_b\Big)\Big)\Big]$ est le tau de Kendall, avec $\widetilde{Z}$ de

même loi que Z. On exploite alors la croissance des h_a pour remarquer que les incréments de Z et de X ont le même signe. En conséquence, $\tau_{a,b} = \mathbb{E}_{X,\widetilde{X}}\Big[\mathrm{sign}\Big(\Big(X_a - \widetilde{X}_a\Big)\Big(X_b - \widetilde{X}_b\Big)\Big)\Big]$. On a alors l'estimateur :

$$\widehat{\tau}_{a,b} = \frac{2}{n(n-1)} \sum_{i<j} \mathrm{sign}\Big(\big(X_a^{(i)} - X_a^{(j)}\big)\Big(X_b^{(i)} - X_b^{(j)}\Big)\Big)\,,$$

dont on déduit une estimation $\widehat{\Lambda}_{a,b}$ de $\Lambda_{a,b}$, et au final du graphe g^* minimal.

5 Tests multiples

5.1 Rappels

Remarque 50 (Cadre) **:**
On considère une famille de lois $(\mathbb{P}_\theta)_{\theta\in\Theta}$. Les données X suivent une loi $\mathbb{P}_\theta$ et on veut tester $\theta \in \Theta_0$ contre $\theta \in \Theta_1$.

Définition 51 (Test) **:**
Un test est une application $\widehat{T} : X \mapsto \widehat{T}(X) \in \{0, 1\}$.

Définition 52 (p-valeur) **:**
Une p-valeur est une variable $\sigma(X)$-mesurable $\widehat{p}$ telle que $\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(\widehat{p} \leq \alpha) \leq \alpha$ pour tout $\alpha \in [0, 1]$.

Une p-valeur $\widehat{p}$ induit un test $\widehat{T} = \mathbb{1}_{\widehat{p}\leq\alpha}$ au seuil α, tel que $\sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(\widehat{T} = 1\right) \leq \alpha$.

Plus généralement, si $\widehat{S}$ est une variable réelle $\sigma(X)$-mesurable, et que $\widehat{T} = \mathbb{1}_{\widehat{S}\geq c}$ pour un seuil $c \in \mathbb{R}$, on peut en déduire une p-valeur.

Lemme 53 :
Soient $F_\theta(s) = \mathbb{P}_\theta\left(\widehat{S} \leq s\right)$ et $T_\theta = 1 - F_\theta$. On pose $\widehat{p}(s) = \sup_{\theta\in\Theta_0} T_\theta(s)$.

Alors $\widehat{p}$ est une p-valeur.

Démonstration. Soit $\theta \in \Theta_0$ fixé. On a une variable aléatoire $U \sim \mathcal{U}([0, 1])$ telle que $\widehat{S} \overset{\mathrm{d}}{=} F_\theta^{-1}(U)$. Alors $\mathbb{P}_\theta\left(\widehat{p}\left(\widehat{S}\right) \leq \alpha\right) \leq \mathbb{P}_\theta\left(T_\theta\left(\widehat{S}\right) \leq \alpha\right) = \mathbb{P}_\theta(U \geq 1 - \alpha) = \alpha$. □

5.2 Tests multiples

On considère désormais m tests $\left(\Theta_0^{(i)}, \Theta_1^{(i)}\right)$, et des p-valeurs $\widehat{p}_i$ associées. On pose $I_0 \subset [m]$ comme l'ensemble des paramètres i pour lesquels $H_0^{(i)}$ est vraie, pour lesquels on a $\theta \in \Theta_0^{(i)}$.

Notre estimateur est ici $\widehat{R} : (p_i) \in [0, 1]^m \to \mathcal{P}([m])$, de sorte que $\widehat{R}(p_1, \ldots, p_m)$ est l'ensemble des indices $i \in [m]$ pour lesquels on rejette $H_0^{(i)}$. On note alors $FP = \left|\widehat{R} \cap I_0\right|$ le nombre de faux positifs, de variables.

Remarque 54 (Bonferroni) **:**

Si $\widehat{R} = \{i \in [m],\, \widehat{p}_i \leq \alpha\}$, alors on a par linéarité $\mathbb{E}[FP] \leq |I_0|\alpha =: m_0\alpha$.

En conséquence, si on utilise $\widehat{R}^{Bonf}$ au seuil $\frac{\alpha}{m}$, on obtient $\mathbb{E}[FP] \leq \frac{m_0}{m}\alpha \leq \alpha$.

Le souci de cette approche est alors que $\left|\widehat{R}\right|$ est aussi très petit.

5.3 Taux de fausses découvertes

On pose $FDR = \mathbb{E}\left[\frac{FR}{\left|\widehat{R}\right|}\mathbb{1}_{\widehat{R}>0}\right]$. On aimerait $\widehat{R}$ qui minimise FDR.

On s'intéresse à $\widehat{R} = \{i \in [m],\, \widehat{p}_i \leq \tau(\widehat{p}_1, \ldots, \widehat{p}_m)\}$, où τ est un seuil fonction des données.

Quitte à trier les valeurs $\widehat{p}_{(1)} \leq \cdots \leq \widehat{p}_{(m)}$ par ordre croissant, on peut alternativement écrire $\widehat{R} = \left\{i \in [m],\, \widehat{p}_i \leq \widehat{p}_{\left(\widehat{k}\right)}\right\}$, avec $\widehat{k}(\widehat{p}_1, \ldots, \widehat{p}_m)$ lui-même aléatoire.

Moralement, on a $\mathbb{E}\left[\frac{FP}{\widehat{R}}\right] \cong \mathbb{E}\left[\frac{m_0 \times \widehat{p}_{(\widehat{k})}}{\widehat{k}}\right] \leq \alpha$. On veut maximiser $\widehat{k}$ en restant sous le seuil α. On considère alors $\widehat{k} = \max\left\{k,\, \widehat{p}_{(k)} \leq \frac{\alpha k}{m}\right\}$.

Définition 55 :

Soit $\beta : [m] \to \mathbb{R}^+$ croissante. La fonction induit une variable $\widehat{k} = \max\left\{k,\, \widehat{p}_{(k)} \leq \frac{\alpha}{m}\beta(k)\right\}$.

Remarque 56 :

Les deux valeurs traditionnelles de β sont $\beta(k) = k$ (Benjamini-Hochberg) et $\beta(k) = \frac{k}{\ln(m)}$ (Bejmanini-Yekuteli).

Théorème 57 :

On a $FDR \leq \frac{m_0}{m}\alpha \sum_{j=1}^{\infty} \frac{\beta(j \wedge m)}{j(j+1)}$.

Démonstration. On a :

$$\begin{aligned} FDR &= \sum_{i \in I_0} \mathbb{E}\left[\mathbb{1}_{\widehat{p}_i \leq \frac{\alpha}{m}\beta\left(\widehat{k}\right)\frac{\mathbb{1}_{\widehat{k}>0}}{\widehat{k}}}\right] \\ &= \sum_{i \in I_0} \sum_{j \geq 1} \tfrac{1}{j(j+1)} \mathbb{E}\left[\mathbb{1}_{\widehat{p}_i \leq \frac{\alpha}{m}\beta\left(\widehat{k}\right)} \mathbb{1}_{j \geq \widehat{k} \geq 1}\right] \\ &\leq \sum_{i \in I_0} \sum_{j \geq 1} \tfrac{1}{j(j+1)} \mathbb{E}\left[\mathbb{1}_{\widehat{p}_i \leq \frac{\alpha}{m}\beta(j \wedge m)} \mathbb{1}_{j \geq \widehat{k} \geq 1}\right] \\ &\leq \sum_{i \in I_0} \sum_{j \geq 1} \tfrac{1}{j(j+1)} \tfrac{\alpha}{m}\beta(j \wedge m)\,, \end{aligned}$$

d'où la majoration voulue. □

Dans certain cas, on peut concevoir des exemples pathologiques où on a égalité ci-dessus, donc cette borne est en général optimale.

Corollaire 58 :
Si $\beta(j) = \gamma j$, alors la majoration devient $\frac{m_0}{m}\alpha\gamma H_m$, d'où le choix $\gamma_m = \frac{1}{H_m} \sim \frac{1}{\ln(m)}$ pour obtenir une majoration par α avec Benjamini-Yekuteli.

Cependant, cette version renormalisée par $\ln(m)$ peut s'avérer nettement moins performante que l'approche plus naïve par Bonferroni sur des petits ensembles d'hypothèses.

5.4 Critères sur Benjamini-Hochberg

On considère ici $\beta(k) = k$. Alors :

$$\begin{aligned} FDR &= \sum_{i\in I_0}\sum_{k=1}^{m} \tfrac{1}{k}\mathbb{E}\left[\mathbb{1}_{\widehat{p}_i \leq \frac{\alpha k}{m}}\mathbb{1}_{\widehat{k}=k}\right] \\ &\leq \sum_{i\in I_0}\sum_{k=1}^{m} \tfrac{1}{k}\tfrac{\alpha k}{m}\mathbb{P}\left(\widehat{k}=k\middle|\widehat{p}_i \leq \tfrac{\alpha k}{m}\right) \\ &\leq \tfrac{\alpha}{m}\sum_{i\in I_0}\sum_{k=1}^{m} \mathbb{P}\left(\widehat{k}\leq k\middle|\widehat{p}_i \leq \tfrac{\alpha k}{m}\right) - \mathbb{P}\left(\widehat{k}\leq k-1\middle|\widehat{p}_i \leq \tfrac{\alpha k}{m}\right) \end{aligned}$$

Si on a $\mathbb{P}\left(\widehat{k}\leq k\middle|\widehat{p}_i \leq \frac{\alpha k}{m}\right) \leq \mathbb{P}\left(\widehat{k}\leq k\middle|\widehat{p}_i \leq \frac{\alpha(k+1)}{m}\right)$, alors chaque somme en k se télescope, et on majore finalement par $\frac{m_0}{m}\alpha \leq \alpha$.

Reste donc à justifier cette inégalité. Cela vient du faire que, si $\widehat{p}_i$ augmente, alors par définition $\widehat{k}$ ne peut que diminuer, et donc $\mathbb{1}_{\widehat{k}\leq k}$ augmente.

Définition 59 (*Positive Regression Dependency on a Subset* (PRDS)) **:**
On dit que $(\widehat{p}_i)$ satisfait PRDS si, pour toute fonction $g : [0,1]^m \to \mathbb{R}^+$ croissante coordonnée par coordonnée, l'application $u \mapsto \mathbb{E}[g(\widehat{p}_1, \ldots, \widehat{p}_m)|\widehat{p}_i \leq u]$ est croissante pour tout $i \in I_0$.

Théorème 60 :
Si on satisfait le critère PRDS, alors en particulier $FDR \leq \alpha\frac{m_0}{m}$ pour Benjamini-Hochberg.

Ce critère est satisfait lorsque les variables sont positivement corrélées, ce qui restreint assez considérablement son champ d'application en pratique.

6 Clustering

Jusqu'ici, on s'est essentiellement intéressé à des échantillons iid, avec des données réparties de façon homogène, relativement à la distribution considérée.

En pratique, on pourrait avoir plusieurs sous-populations distinctes. On suppose donc qu'il existe une *petite* famille de paramètres $\{(\theta_k, \Lambda_k),\ 1 \leq k \leq K\}$ et qu'on considère un échantillon indépendant $X_i \sim \mathcal{N}(\mu_i, \Sigma_i)$ avec $(\mu_i, \Sigma_i) = (\theta_k, \Lambda_k)$ pour un certain k.

Autrement dit, on peut en théorie partitionner notre échantillon en K sous-populations iid et indépendantes entre elles. Cependant, on ne connait pas *a priori* cette partition $[n] = \bigsqcup_{k=1}^{K} G_k$.

Une approche naïve est de faire du clustering par agglomérats, en partant de n composantes, et en reliant les deux composantes les plus proches par une arête, jusqu'à obtenir le nombre voulu de composantes. Cette approche n'est pas forcément adaptée à la géométrie du problème, mais surtout elle est purement locale.

Pour une approche plus globale, on peut considérer le maximum de vraisemblance, mais elle fait appel à l'inverse de matrices de taille p, ainsi qu'à un minimum parmi toutes les k-partitions de $[n]$, ce qui rend les calculs absurdement lourds et très instables.

Pour simplifier l'estimation, on suppose généralement qu'on a $\Lambda_k = \sigma^2 I$. Dans ce cas, le maximum de vraisemblance devient K-means :

$$\widehat{G}_{K-means} \in \operatorname*{argmin}_{G \text{ partitionne } [n]} \sum_{k=1}^{K} \min_{\theta_k \in \mathbb{R}^p} \left(\sum_{i \in G_k} \|X_i - \theta_k\|^2 \right) = \operatorname*{argmin}_{G} \sum_{k=1}^{K} \sum_{i \in G_k} \left\| X_i - \overline{X}_{G_k} \right\|^2 .$$

Si on développe le produit scalaire, par définition de $\overline{X}_{G_k}$, on obtient :

$$\sum_{i \in G_k} \left\| X_i - \overline{X}_{G_k} \right\| = \sum_{i \in G_k} \left\langle X_i - \overline{X}_{G_k}, X_i \right\rangle = \frac{1}{2|G_k|} \sum_{i,j \in G_k} \|X_i - X_j\|^2 .$$

En passant à la somme sur k, on a donc :

$$\mathbb{E}[crit_K(G)] = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i,j \in G_k} \mathbb{E}\left[\|X_i - X_j\|^2 \right] = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i,j \in G_k} \|\mu_i - \mu_j\|^2 + \sum_{i=1}^{n} \gamma_i + \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i \in G_k} \gamma_i ,$$

où $\gamma_i = \operatorname{Tr}(\operatorname{Cov}(X_i))$. Le terme de gauche est minimisé par le bon choix de partition, qui l'annule. Le terme central est constant, ne dépend plus de G. Enfin, pour le terme de droite, il nous faudrait en général une façon d'estimer ces variances sans avoir pu faire le clustering en premier lieu. Sous l'hypothèse simplificatrice que tous les X_i ont la même covariance, le terme devient constant.

Pour le calcul exact de $\widehat{G}_{K-means}$, on peut montrer que tout algorithme d'approximation est NP-dur. On a donc plutôt assouplir le problème par convexification. Si on considère $B_{i,j}(G) = \sum_{k=1}^{K} \frac{\mathbb{1}_{i,j \in G_k}}{|G_k|}$, et $D_{i,j} = \|X_i - X_j\|^2$. Dans ce cas, $\widehat{G}_{K-means} = \operatorname*{argmin}_{G} \frac{1}{2} \langle B(G), D \rangle$.

Lemme 61 :

On peut montrer que $\{B(G),\, G \text{ partitionne } [n]\}$ est égal à :

$$\left\{B \in S_n^+,\, B^2 = B,\, \mathrm{Tr}(B) = K,\, B \text{ stochastique à droite}\right\},$$

autrement dit B est une projection sur un espace de dimension K.

La seule condition non-linéaire dans le terme de droite est $B^2 = B$. C'est ce qui rend le problème difficile. Si on retire ce critère, on peut alors calculer $\widehat{B}^{SDP} \in \mathrm{argmin}\langle B, D\rangle_F$ parmi les matrices symétriques positives, stochastiques à droite, de trace K. Malheureusement, outre le problème de quoi faire de cette estimation de $B(D)$, dès que n est grand l'algorithme est très long à converger.

En pratique, on utilise une autre approche. A θ fixé, optimiser G est facile, il suffit de prendre les ensembles de Voronoi. A G fixé, optimiser θ est facile, il suffit de prendre la moyenne des points de chaque cluster. L'algorithme de Lloyd consiste à itérer ces deux phases, de façon gloutonne, en partant d'une partition $\widehat{G}^0$ quelconque.

On a un terme de biais dans Lloyd, même lorsque tous les γ_i sont égaux. En supposant qu'ils sont tous égaux, on peut estimer $\widehat{\gamma} = \frac{1}{2}\min_{i\neq j}\|X_i - X_j\|^2$, et débiaiser l'algorithme de Lloyd.

En admettant que cet algorithme converge vers un minimum local, cette limite n'est pas forcément le minimum global. Le résultat de l'algorithme dépend fortement de l'initialisation, il faut donc également avoir une heuristique pour calculer $\widehat{G}^0$.

Pour ce faire, on va utiliser des méthodes de clustering spectrales. On écrit :

$$\mathbb{X} = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} = \begin{pmatrix} \mu_1^T \\ \vdots \\ \mu_n^T \end{pmatrix} + E = A \begin{pmatrix} \theta_1^T \\ \vdots \\ \theta_K^T \end{pmatrix} + E\,,$$

où E est le vecteur d'erreurs, et $A_{i,k} = \mathbb{1}_{i\in G_k}$. La matrice A est de rang K.

On a $\mathbb{X}\mathbb{X}^T = A\Theta\Theta^T A^T + EE^T + A\Theta E^T + E\Theta^T A^T$.

6.1 Bornes de récupération

On se place dans le cadre de travail le plus simple : on a deux groupes, centrés en $\pm\theta$. Autrement dit on a $X_i = z_i\theta + E_i$, avec $z_i = \pm 1$ et $E_i \sim \mathcal{N}(0, \sigma^2 I_p)$.

Notre métrique à optimiser ici est $recov(\widehat{z}) = \frac{\|\widehat{z}-z\|_0}{n}$, la proportion de points bien assignés. Pour éviter le problème du choix de signe non canonique, on peut remplacer cette métrique par son minimum entre $\widehat{z}$ et $-\widehat{z}$.

À θ fixé, l'estimateur de Bayes est $h^*(x) = \text{sign}(\mathbb{E}[Z|X=x]) = \text{sign}(\langle\theta, x\rangle)$. Pour classifier les points, il faut donc déjà pouvoir estimer θ. Si on a déjà un estimateur z sur un échantillon de taille n, on remplace θ par $\frac{1}{n}\sum_{i=1}^{n} z_i X_i$.

Soient (X_i) un échantillon bien classifié par (z_i), et X le nouveau point reçu. On s'intéresse à :

$$\mathbb{P}(h^*(X) \neq Z) = \mathbb{P}\left(\left\langle \frac{1}{n}\sum_{i=1}^{n} z_i\theta + z_i E_i, ZX \right\rangle < 0\right).$$

Quitte à noter $z_i E_i = \varepsilon_i$, et $ZX = \theta + \varepsilon'$, on peut réécrire cette probabilité :

$$\mathbb{P}\left(\|\theta\|_2^2 + \left\langle \theta, \varepsilon' + \frac{\varepsilon}{\sqrt{n}} \right\rangle + \frac{\langle\varepsilon, \varepsilon'\rangle}{\sqrt{n}} < 0\right).$$

Le produit scalaire de gauche est asymptotiquement de loi $\|\theta\|\sigma\mathcal{N}(0,1)$, et celui de droite $\sqrt{p}\sigma^2\mathcal{N}(0,1)$ indépendant. Pour de grandes valeurs de n, on est donc proche du cas suivant :

$$\mathbb{P}\left(\|\theta\|^2 + \sigma\sqrt{\|\theta\|^2 + \frac{p}{n}\sigma^2}N < 0\right)$$

avec $N \sim \mathcal{N}(0,1)$, et alors on a la majoration :

$$\mathbb{P}\left(N < -\frac{\|\theta\|^2}{\sigma\sqrt{\|\theta\|^2 + \frac{p}{n}\sigma^2}}\right) \leq \exp\left(-\frac{s^2}{2}\right),$$

où $s^2 := \frac{\|\theta\|^4}{\sigma^2\left(\|\theta\|^2 + \frac{p}{n}\sigma^2\right)}$.

À partir de cette borne, on aimerait dans l'idéal obtenir une borne du type $recov(\widehat{z}) \leq e^{-cs^2}$.

Retournons au cas spectral de la semaine dernière. On a ici $\mathbb{X}\mathbb{X}^T = \|\theta\|^2 zz^T + EE^T + E\theta z^T + z\theta^T E^T$. En passant à la moyenne, les termes de droite disparaissent, et les seuls termes de EE^T qui ne sont pas des produits de variables indépendantes sont le long de la diagonale. On a donc $\mathbb{E}\left[\mathbb{X}\mathbb{X}^T\right] = \|\theta\|^2 zz^T + p\sigma^2 I_n$.

Soit $\widehat{v}$ un vecteur propre normalisé, associé à la valeur propre maximale de $\mathbb{X}\mathbb{X}^T$. On s'intéresse à $\widehat{z} = \text{sign}(\widehat{v})$.

Théorème 62 :

Il existe $C > 0$ telle que, avec probabilité au moins $1 - 3e^{-\frac{n}{2}}$, on a $recov(\widehat{z}) \leq \frac{C}{s^2}$.

Idées. On peut écrire $\mathbb{X}\mathbb{X}^T = \mathbb{E}\big[\mathbb{X}\mathbb{X}^T\big] + W$, et le vecteur propre maximal de $\mathbb{E}\big[\mathbb{X}\mathbb{X}^T\big]$ est $v = \frac{z}{\sqrt{n}}$.

On utilise pour cela des bornes de perturbation. Avec la borne de Davis-Kahan, on a l'inégalité $\big\|\widehat{v}\widehat{v}^T - vv^T\big\|_F^2 \leq 8\frac{\|W\|^2}{(\lambda_1 - \lambda_2)}$, en lien avec le trou spectral, où $\|W\|$ est la norme d'opérateur.

En outre, on peut montrer que $recov(\widehat{z}) \leq \big\|\widehat{v}\widehat{v}^T - vv^T\big\|_F^2$.

En combinant ces résultats, on s'est ramené d'un problème de majoration sur $recov(\widehat{z})$ à une majoration sur $\|W\|$.

On montre alors qu'avec probabilité au moins $1 - 2e^{-\frac{n}{2}}$, on a $\|W\|_{0,p} \leq C\frac{n\|\theta\|^2}{s}$. Pour ce faire, on revient à l'écriture $W = \big(EE^T - p\sigma^2 I_n\big) + E\theta z^T + z\theta^T E^T$.

Pour la parenthèse de gauche, on va majorer sa norme par $C\big(\sqrt{np} + n\big)$ avec probabilité $1 - 2e^{-\frac{n}{2}}$.

Dans le cas d'une matrice symétrique, on a $\|A\| = \sup_{\|u\|_2 = 1} |\langle Au, u\rangle|$. Ici, pour W, on a en particulier $\langle Wu, u\rangle = \big\|E^T u\big\|_2^2 - p = \|\varepsilon\|_2^2 - p$ avec $\varepsilon \sim \mathcal{N}(0, I_p)$. En utilisant la borne de Hanson-Wright, avec les matrices $\Sigma = \pm I_p$, on a avec grande probabilité :

$$\mathbb{P}\Big(|\langle Au, u\rangle| \geq C\Big(\sqrt{pL} \vee L\Big)\Big) \leq 2e^{-L}\,.$$

Pour passer de cette borne au supremum sur la sphère, on a besoin d'utiliser $\mathcal{N}_\varepsilon$ un *ε-net*, une famille finie de points de la sphère telle que pour tout $x \in \partial B(0,1)$, $d(x, \mathcal{N}_\varepsilon) \leq \varepsilon$. Dans ce cas, on a assez naturellement $\|A\| \leq \frac{1}{1-2\varepsilon} \max_{y \in \mathcal{N}_\varepsilon} |\langle Ay, y\rangle|$.

Or on peut obtenir un $\mathcal{N}_\varepsilon$ de cardinal inférieur à $\big(1 + \frac{2}{\varepsilon}\big)^n$ en dimension n. Pour ce faire, on prend une suite de points (x_n) telle que $x_{n+1} \notin \bigcup_{i=1}^{n} B(x_i, \varepsilon)$. Par compacité, on extrait une sous-suite finie $\mathcal{N}_\varepsilon$ telle que la famille de boules de rayons ε recouvre la sphère. Par construction, les boules de rayons $\frac{\varepsilon}{2}$ sont disjointes, donc $|\mathcal{N}_\varepsilon| \times \mathrm{Vol}\big(B\big(0, \frac{\varepsilon}{2}\big)\big) \leq \mathrm{Vol}\big(B\big(0, 1 + \frac{\varepsilon}{2}\big)\big)$, dont on déduit la borne souhaitée.

Par sous-additivité, on a donc $\mathbb{P}\big(\big\|EE^T - pI_n\big\| \geq \frac{C}{1-2\varepsilon}\big(\sqrt{pL'} \vee L'\big)\big) \leq 2e^{-L'}$ avec la nouvelle constante $L' = L + n\ln\big(1 + \frac{2}{\varepsilon}\big)$.

En se ramenant à des gaussiennes, on a en outre $\mathbb{P}\big(\big\|E\theta z^T\big\| \geq C\|\theta\|_2 n\big) \leq e^{-\frac{n}{2}}$.

On peut finalement assembler ces bornes pour conclure sur le résultat annoncé. □

Revenons à l'algorithme de Lloyd. Dans ce cas, cela revient à itérer l'opérateur $\widehat{z}^{t+1} = \mathrm{sign}\big(\mathbb{X}\mathbb{X}^T\widehat{z}^t\big)$ en partant de $\widehat{z}^0$. La différence fondamentale avec le cas spectral vu ici est que, dans Lloyd, on extrait les signes à chaque étape, là où on extrait uniquement le signe après avoir appliqué l'opérateur linéaire une infinité de fois dans le cas spectral.

Lorsque $\widehat{z}^0$ est issu du cas spectral, on peut prouver que, pour Lloyd, $recov\left(\widehat{z}^{Lloyd}\right) \leq e^{-cs^2}$. En outre, la convergence des itérées de Lloyd vers cette borne est assez rapide, il suffit d'un $O(\ln(n))$ itérations pour converger.