Apprentissage en grande domension

March 15, 2017

$$\min_{\beta \in \mathbb{R}} f(\beta) \tag{1}$$

Conditions: f convexe:

$$f(y) >= f(x) + \nabla f(x)^T (y - x) \tag{2}$$

Definition 1:

$$\forall \theta \in [0, 1] \tag{3}$$

Def 3 M

Def 4 Lipschizsienne

$$\forall x, y ||f(x) - f(y)||_2 <= L||x - y||_2 \tag{4}$$

Def 5 contractant

$$L Lipschitz avec 0 <= L < 1 \tag{5}$$

Them 1 Thm point fixe: f est $\alpha-$contractant,

$$\exists x^* telque f^* = f(x^*) \tag{6}$$

La suite definie par $x_{n+1} = f(x_n)$ converge vers x^* et vérifie

$$||x_n - x^*||_2 <= \frac{\alpha^n}{1 - \alpha} ||x_0 - x_1||_2 \tag{7}$$

Gradient Algo

Prop 5 Gradient monotone f diff est convexe, si et seulement si

$$\begin{aligned} &(\nabla f(x) - \nabla f(y))^T (x - y) >= 0 \\ &= \nabla f(x) f - \text{consistante} \end{aligned} \tag{8}$$

PREUVE 1.$\Rightarrow$:

$$f(y) >= f(x) + \nabla f(x)^T (y - x) \tag{9}$$

$$f(x) >= f(y) + \nabla f(y)^T (x - y) \tag{10}$$

$$-f(x) - f(y) < -f(x) - f(y) + \nabla f(x)^T (x - y) - \nabla f(y)^T (x - y) \tag{11}$$

$$(\nabla f(x) - \nabla f(y))^T (x - y) >= 0 \tag{12}$$

2. $\Leftarrow$: On introduit une fonction Φ:

$$\Phi(t) = f(x + t(y - x)) \tag{13}$$

$$\Phi'(t) = \nabla f(x + t(y - x))^T (y - x) \tag{14}$$

Comme ∇f est monotone

$$\Phi'(t) >= \Phi'(0), t >= 0 \tag{15}$$

$$f(y) - \Phi(1) = \Phi(0) + \int_0^1 \Phi'(t)dt \tag{16}$$

$$f(y) >= \Phi(0) + \Phi'(0) = f(x) + \nabla f(x)^T (y - x) \tag{17}$$

Theorème Boîte quadratique supérieure

$$f \sim L^1, \nabla f est L - lipschitz \tag{18}$$

Alors

$$g(x) = \frac{L}{2} x^T x - f(x) est convexe \tag{19}$$

$$f(y) <= \nabla <= \nabla f(x)^T (y - x) + \frac{L}{2} ||x - y||_2^2 \tag{20}$$

1. ∇f Lipschitz

$$||\nabla f(y) - \nabla f(x)||_2 <= L||y - x||_2 \tag{21}$$

2.

$$\begin{aligned} (\nabla f(y) - \nabla f(x))^T (y - x) &<= ||\nabla f(y) - \nabla f(x)||_2 ||y - x||_2 \\ &<= L||y - x||_2^2 \end{aligned} \tag{22}$$

$$\nabla g(x) = Lx - \nabla f \tag{23}$$

$$\begin{aligned} &(\nabla g(x) - \nabla g(y))^T (x - y) \\ =&(Lx - \nabla f(x) - Ly + \nabla f(y))^T (x - y) \\ =& - (\nabla f(y) - \nabla f(x))^T (y - x) + L||x - y||_2^2 \\ >=&0 \end{aligned} \tag{24}$$

$$y = x - t\nabla f(x) \tag{25}$$

$$f(x - t\nabla f(x)) <= f(x) + t(1 - \frac{Lt}{2})||\nabla f(x)||_2^2 \tag{26}$$

choix de t tel que $0 <= t < \frac{1}{2}$

$$x^+ = x - t\nabla f(x) \tag{27}$$

$$\begin{aligned} f(x^+) <=& f(x) + f(1 - \frac{Lt}{2})||\nabla f(x)||_2^2 \\ <=& f(x) - \frac{t}{2}||\nabla f(x)||_2^2 \end{aligned}$$

$$\begin{aligned} <= f^* + \nabla f(x)^T (x - x^*) - \frac{t}{2}||\nabla f(x)||^2 \\ =&f^* + \frac{1}{2t}(||x - x^*||_2^2 - ||x - x^* - t\nabla f(x)||_2^2) \\ =&f^* + \frac{1}{2t}(||x - x^*||_2^2 - ||x^+ - x^*||_2^2) \end{aligned} \tag{28}$$

$$\begin{aligned}\sum_{k=1}^{N}(f(x_k)-k^*) <= &\frac{1}{2t}\sum_{k=1}^{N}(||x_{k-1}-x^*||_2^2-||x_k-x^*||_2^2)\\ =&\frac{1}{2t}(||x_0-x^*||_2^2-||x_N-x^*||_2^2)\\ <= \frac{1}{2t}||x_0-x^*||_2^2\end{aligned} \tag{29}$$

Prop: Quand f est differenciable

$$f(y) >= f(x) + \nabla f(x)^T(y-x) \tag{30}$$

Definition: sous gradient g est un sous gradient de f en x, ssi

$$\forall y, f(y) >= f(x) + g^T(y-x) \tag{31}$$

Definition: sous differentielle f convexe, on definit la sous differentielle de f en x comme

$$\partial f(x) = \{g|\forall y, f(y) >= f(x) + g^T(y-x)\} \tag{32}$$

Theoreme 3:

$$x^* = argminf \Leftrightarrow 0 \in \partial f(x^*) \tag{33}$$

Si $0 \in \partial f(x^*)$, alors

$$\forall y, f(y) >= f(x^*) + 0^T(y-x^*) \Leftrightarrow x^= argminf \tag{34}$$

Prop 7: linéarité non négative f_1 et f_2 convexes, $\alpha_1, \alpha_2 >= 0$

$$f >= \partial(\alpha_1 f_1 + \alpha_2 f_2)(x) = \alpha_1 \partial f_1(x) + \alpha_2 \partial f_x(x) \tag{35}$$

+ addition d'ensemble

$$E + F = \{e + favece \in E, f \in F\} \tag{36}$$

Prop 8: combinaison affine: Si $h(x) = f(Ax+b)$, alors

$$\partial h(x) = A^T \partial f(Ax+b) \tag{37}$$

f est une fonction G-Lipschitzienne
ALGO: Méthode du "sous-gradient"

$$x_k \leftarrow x_{k-1} - t_k g_{k-1} \tag{38}$$

ou

$$g_{k-1} \in \partial f(x_k - 1) \tag{39}$$

Trois possibilité pour t_k

1. $t_k = t$

2. "Longueur constante" $t_k ||g_{k-1}||_2 estconstante$

3.

$$t_k \rightarrow_{k \rightarrow +\infty} 0 \tag{40}$$

$$\sum_{k=1}^{+\infty} = +\infty \tag{41}$$

$$\sum_{k=1}^{+\infty} t_k^2 = \text{limite finie} \tag{42}$$

Theoreme: f convexe et non differentielle f est G-Lipschitzienne $\Leftrightarrow ||g||_2 <= G, \forall g \in \partial f(x)$

Preuve: $\Leftarrow$

On suppose $\forall x, \forall g \in \partial f(x)$

$$||g||_2 <= G \tag{43}$$

Soit $x(g_x)$ et $y(g_y)$

$$g_x^T(x-y) >= f(x) - f(y) >= g_y^T(x-y) \tag{44}$$

$$G||x-y||_2 >= f(x) - f(y) >= -G||x-y||_2 \tag{45}$$

$$\forall x, y, ||f(x) - f(y)|| <= G||x-y||_2 \tag{46}$$

$\Rightarrow \exists g$ tel que $||g||_2 > G$

$$y = x + \frac{g}{||g||_2} \tag{47}$$

$$f(y) >= f(x) + g^T(y-x) = f(x) + ||g||_2 > f(x) + G \tag{48}$$

Pas possible car f est G-Lipschitzienne

Attention: La méthode du sous-gradient n'est pas une méthode de descente.

$$x^+ = tg \tag{49}$$

g sous-gradient de f en x.

$$\begin{aligned} ||x^+ - x^*||_2^2 = ||x - tg - x^*||_2^2 & \\ &= ||x - x^*||_2^2 + t^2||g||_2^2 - 2tg^T(x - x^*) \\ &<= ||x - x^*||_2^2 + t^2||g||_2^2 - 2t(f(x) - f^*) \end{aligned} \tag{50}$$

Pour une iteration k:

$$2t_k(f(x_{k-1}) - f^*) < ||x_{k-1} - x^*||_2^2 - ||x_k - x^*||_2^2 + t_k^2||g_{k-1}||_2^2 \tag{51}$$

en sommant les inégalités

$$\begin{aligned}
2(\sum_{k=1}^{N} t_k)(f_{best}^{(}N) - f^*) &<= ||x_0 - x^*||_2^2 - ||x_N - x^*||_2^2 + \sum_{k=1}^{N} t_k^2 ||g_{k-1}||_2^2 \\
&<= ||x_0 - x^*||_2^2 + \sum_{k=1}^{N} t_k^2 ||g_{k-1}||_2^2
\end{aligned} \tag{52}$$

1. $t_k = t$

$$f_{best}^{(N)} - f^* <= \frac{||x_0 - x^*||_2^2}{2Nt} + \frac{G^2 t}{2} \tag{53}$$

2. $t_k ||g_{k-1}||_2 = s$

$$f_{best}^{(N)} - f^* <= \frac{G||x_0 - x^*||_2^2}{2Ns} + \frac{Gs}{2} \tag{54}$$

3. $t_k \to 0, \sum t_k \to +\infty, \sum t_k^2$ converge

$$f_{best}^{(N)} - f^* <= \frac{||x_0 - x^*||_2^2 + \sigma^2 \sum t_k^2}{2 \sum t_k} \tag{55}$$

Conclusion: La méthode du sous gradient n'est pas facile à paramétrer pour obtenir sa convergence.

Exercise:

$$f(\beta) = ||X\beta - y||_2^2 + \lambda ||\beta||_1 \tag{56}$$

$$\partial f(\beta) = X^T (X\beta - y) + \lambda \partial_{||\dot{|}|_1}(\beta) \tag{57}$$

$$[\partial_{||\dot{|}|_1}(\beta)] = \begin{cases} sign(\beta_i) & \text{si}\beta_i \neq 0 \\ [-1, 1] & \text{si}\beta_i = 0 \end{cases} \tag{58}$$

Definition Operateur proximal

$$prox_f(x) = argmin_u \{ f(u) + \frac{1}{2} ||u - x||_2^2 \} \tag{59}$$

f convexe "semi-continue inférieurement"(sci). alors, $prox_f(x)$ existe et est unique.

Theoreme Caractérisation par le sous-gradient

$$u = prox_f(x) \Leftrightarrow x - u \in \partial f(u) \tag{60}$$

Preuve:

$$\begin{aligned}
u = prox_f(x) &\Leftrightarrow u = argmin\{ f(u) + \frac{1}{2} ||u - x||_2^2 \} \\
&\Leftrightarrow 0 \in \partial g(u) \\
&\Leftrightarrow 0 \in \partial g_1(u) + \partial g_2(u) \\
&\Leftrightarrow 0 \in \partial f(u) + (u - x) \Leftrightarrow x - u \in \partial f(u)
\end{aligned} \tag{61}$$

$$g(y) = g_1(y) + g_2(y) = f(y) + \frac{1}{2}||y - x||_2^2 \tag{62}$$

Algorithme du gradient proximal

$$0 \in \partial f(x^*) \Leftrightarrow x^* = argmin_x f(x) \tag{63}$$

$$\partial(f_1 + f_2) = \partial f_1 + \partial f_2 \tag{64}$$

Si f est différentielle en x, alors

$$\partial f(x) = \nabla f(x) \tag{65}$$

Norme euclidienne

$$f(x) = ||x||_2 \tag{66}$$

$$prox_{tf}(x) = \left\{ \begin{array}{cc} (1 - \frac{t}{||x||_2})x & , ||x||_2 >= t \\ 0 & , sinon \end{array} \right. \tag{67}$$

Multiplication par un scalaire ¿0

$$f(x) = \lambda g(x/\lambda) \tag{68}$$

$$prox_f(x) = \lambda prox_{\frac{1}{\lambda}g}(\frac{x}{\lambda}) \tag{69}$$

Somme séparable (Group LASSO)

$$f([x, y] = g(x) + h(y) \tag{70}$$

$$prox_f([x, y]) = [prox_g(x), prox_h(y)] \tag{71}$$

Norme l_1

$$f(x) = ||x||_1 \tag{72}$$

$$[prox_f(x)]_i \left\{ \begin{array}{cc} x_i - 1 & \text{si} x_i >= 1 \\ 0 & \text{si} |x_i| < 1 \\ x_i + 1 & \text{si} x_i <= -1 \end{array} \right. \tag{73}$$

Numériquement

$$proxl_1(x) = sign(x) \times pmax(abs(x) - 1, x) \tag{74}$$

$$min_\beta f(\beta) = min_\beta\{g(\beta) + h(\beta)\} \tag{75}$$

Algorithme du gradient proximal g convexe et differentiable, ∇g est L-Lipschitzienne

h convexe et non-differentiable (sci pour avoir $prox_{l_2}(x)$)

Exercise

$$f(\beta) = ||X\beta - y||_2^2 + \lambda||\beta||_1 \tag{76}$$

Algorithme:

$$x_k \leftarrow prox_{t_k h}(x_{k-1-t_k \nabla g(x_{k-1})}) \tag{77}$$

$$f^* = f(x^*) \text{ fini} \tag{78}$$

$$t_k = \frac{1}{L}, (0 <= t_k < \frac{1}{L}) \tag{79}$$

Gradient Map

$$G_t(x) = \frac{1}{t}(x - prox_{tl_2}(x - t\nabla g(x))) \tag{80}$$

Pourquoi?

$$x^+ = x - tG_t(x) \tag{81}$$

Attention:

- $G_t(x)$ n'est pas un gradient pour g, n'est pas un sous-gradient pour h ou pour f
- $G_t(x^*) = 0$ ssi $x^* = argminf$

Borne Quadratique Supérieure (BQS)

$$g(y) <= g(x) + \nabla g(x)^T(y - x) + \frac{L}{2}||y - x||_2^2 \tag{82}$$

Pour

$$y(= x^+) = x - tG_t(x) \tag{83}$$

$$\begin{aligned} g(x - tG_t(x)) <=& g(x) - t\nabla g(x)^T G_t(x) + \frac{L}{2}t^2||G_t(x)||_2^2 \\ <= g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2}||G_t(x)||_2^2 \end{aligned} \tag{84}$$

Théorème: L'inégalité précédente nous permet de montrer

$$f(x - tG_t(x)) <= f(z) + G_t(x)^T(x - z) - \frac{t}{2}||G_t(x)||_2^2 \tag{85}$$

$$\begin{aligned} f(x - tG_t(x)) <=& g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2}||G_t(x)||_2^2 + h(x - tG_t(x)) \\ <=& g(z) + \nabla g(z)^T(x - z) - t\nabla g(x)^T G_t(x) + \frac{t}{2}||G_t(x)||_2^2 + h(z) + v^T(x - z - tG_t(x)) \\ =& f(z) + G_t(x)^T(x - z) - \frac{t}{2}||G_t(x)||_2^2 \end{aligned} \tag{86}$$

Pour

$$z = x \tag{87}$$

on a

$$f(x^+) <= f(x) - \frac{t}{2}||G_t(x)||_2^2 \tag{88}$$

$$f(x^+) \to f(x_k) \tag{89}$$

Donc, on a une méthode de descente !

Pour $z = x^*$

$$\begin{aligned} f(x^*) - f^* <=& G_t(x)^T(x - x^*) - \frac{t}{2}||G_t(x)||_2^2 \\ =& \frac{1}{2t}(||x - x^*||_2^2 - ||x - x^* - tG_t(x)||_2^2) \\ =& \frac{1}{2t}(||x - x^*||_2^2 - ||x^+ - x^*||_2^2) \end{aligned} \tag{90}$$

$$f(x_N) - f^* <= \frac{1}{2Nt}||x_0 - x^*||_2^2 \tag{91}$$

$$[prox_{t||||_1}](x) = \begin{cases} x_i - t & \text{si} x_i >= t \\ 0 & \text{si} |x_i| < t \\ x_i + t & \text{si} x_i <= t \end{cases} \tag{92}$$

Fast Proximal gradient algorithm

Convexe & differentielle

$$f(y) >= f(x) + \nabla f(x)^T(y - x) \tag{93}$$

Sous-gradient — sous differentielle

$$\partial f(x) = \{g | g^T(x - y) <= f(y) - f(x)\} \tag{94}$$

Prox.

$$prox_f(x) = argmin_\mu\{f(\mu) + \frac{1}{2}||x - \mu||_2^2\} \tag{95}$$

$$x - u \in \partial f(u) \Leftrightarrow u = prox_f(x) \tag{96}$$

$$\min f(\beta) = g(\beta) + h(\beta) \tag{97}$$

∇g L-Lipschitzienne $prox_{th}$ convexe

FISTA: (n'est pas une méthode de descente)

$$y = x_{k-1} + \frac{k-2}{k+1}(x_{k-1} - x_{k-2}) \tag{98}$$

$$x_k = prox_{t_k h}(y - t_k \nabla g(y)) \tag{99}$$

$$t_k = \frac{1}{L}\text{constant} \tag{100}$$

Reformulation

$$\theta_k = \frac{2}{k+1} \tag{101}$$

v_k tel que $v_0 = x_0$ et $\forall k >= 1$

$$\begin{cases} y = (1-\theta_k)x_{k-1} + \theta_k v_{k-1} \\ x_k = prox_{th}(y - t_k \nabla g(y)) \\ v_k = x_{k-1} + \frac{1}{\theta_k}(x_k - x_{k-1}) \end{cases} \tag{102}$$

Inégalité

$$\forall k >= 2, \frac{1-\theta_k}{\theta_k} <= \frac{1}{\theta_{k-1}^2} \tag{103}$$

BQS(g)

$$g(u) <= g(z) + \nabla g^T(z)(u-z) + \frac{L}{2}||u-z||_2^2 \tag{104}$$

$BQS(h)$

$$u = prox_{th}(w) \tag{105}$$

alors

$$\forall z, h(u) <= h(z) + \frac{1}{t}(v-u)^T(u-z) \tag{106}$$

1.

$$g(x^+) <= g(y) + \nabla g^T(y)(x^+ - y) + \frac{1}{2t}||x^+ - y||_2^2 \tag{107}$$

2.

$$\begin{aligned} h(x^+) &<= h(z) + \frac{1}{t}(y - t\nabla g(y) x^+)^T(x^+ - z) \\ &= h(z) + \nabla g(y)^T(z - x^+) + \frac{1}{t}(x^+ - y)^T(z - x^+) \end{aligned} \tag{108}$$

1+2:

$$\begin{aligned} f(x^+) &= g(x^+) + h(x^+) \\ &<= g(y) + h(z) + \nabla g(y)^T(x^+ - y + z - x^+) + \frac{1}{2t}||x^+ - y||_2^2 + \frac{1}{t}(x^+ - y)^T(z - x^+) \\ &<= f(z) + \frac{1}{2t}||x^+ - y||_2^2 + \frac{1}{t}(x^+ - y)^T(z - x^+) \end{aligned} \tag{109}$$

$$\begin{aligned} & f(x^+) - f^* - (1-\theta)(f(x) - f^*) \\ & <= \frac{\theta^2}{2t}(||v - x^*||_2^2) - ||v^+ - x^*||_2^2 \\ \Leftrightarrow \frac{t}{\theta^2}(f(x) - f^* + \frac{1}{2}||v_1 - x^*||_2^2 & <= \frac{1-\theta_1^2}{\theta_1^2}(f(z) - f^*) + \frac{1}{2}||v - x^*||_2^2 \end{aligned} \tag{110}$$

Comme

$$\frac{1-\theta_1}{\theta_1^2} <= \frac{1}{\theta_{i-1}^2} \tag{111}$$

Conclusion

$$\frac{t}{\theta_k^2}(f(x_k)-f^*)-\frac{1}{2}||v_1-x^*||_2^2 <= \frac{(1-\theta_1)^t}{\theta_1^2}(f(x_0)-f^*)+\frac{1}{2}||v_0-x^+||_2^2 \tag{112}$$

Ainsi

$$\frac{t}{\theta_k^2}f(x_k)-f^* <= \frac{(1-\theta_1)^t}{\theta_1^2}(f(x_0-f^*))+\frac{1}{2}||v_k-x^*||_2^2-\frac{1}{2}||v_0-x^*||_2^2 \tag{113}$$

$$f(x_k)-f^* <= \frac{2L}{(k+1)^2}||x_0-x^*||_2^2 \tag{114}$$

Travaux pratiques: Analyse en Composantes Principales parcimonieuse

1. Equation normale

$$\begin{aligned}
f(v) =&\frac{1}{2}||A-\delta vv^T||_F^2\\
=&\frac{1}{2}tr((A-\delta vv^T)(A-\delta vv^T))\\
=&tr(A^TA-\delta A^Tvv^T-\delta vv^TA+\delta^2vv^Tvv^T)\\
=&\frac{1}{2}tr(A^TA)-\frac{1}{2}tr(\delta A^Tvv^T)-\frac{1}{2}tr(\delta vv^TA)+\frac{1}{2}tr(\delta^2vv^Tvv^T)\\
=&\frac{1}{2}tr(A^TA)-\delta v^TAv+\frac{1}{2}\delta^2(v^Tv)
\end{aligned} \tag{115}$$

$$\delta = v^TAv/(v^Tv)^2 \tag{116}$$

$$Av = \frac{\alpha+\delta^2}{\delta}v \tag{117}$$

Tel que $v^Tv=1, Av=\delta v$ δ est valeur propre de A associée à v

$$f(v) = (v^Tv)^2 \tag{118}$$

$$\nabla f(v) = 4(v^Tv)v \tag{119}$$

$$\begin{aligned}
\nabla L(v) =&-2\delta Av+2\delta^2(v^Tv)v+2\alpha v\\
=&0 \Leftrightarrow \delta Av = (\delta^2(v^Tv)+\alpha)v
\end{aligned} \tag{120}$$

$$f(v,\delta) = \frac{1}{2}||A-\delta vv^T||_F^2 \tag{121}$$

$$A = V\Delta V^T \tag{122}$$

Avec

$$\Delta = diag(\delta_1,\dots,\delta_n) \tag{123}$$

δ, v qui sont solution de $\min f$

$$A-\delta vv^T = Vdiag(0,\delta_2,\dots,\delta_n)V^T = B \tag{124}$$

$$tr(B^T B) = \sum_{k=2}^{m} \delta_k^2 \tag{125}$$

$$\delta_1 = \delta_{\max} \tag{126}$$

ACP

$$v^{k+1} = normalize(Av^{(k)}) \tag{127}$$

ACP l_1

$$V^{(k+1)} = normalize(prox_{\lambda|||}(Av^{(k)})) \tag{128}$$

1 Les modèle graphique gaussien

Soit $X \sim \mathcal{N}(\mu, \Sigma)$, on suppose Σ inversible, de dentité

$$f_\alpha(x) = (2\Pi)^{-P/2}|\Sigma|^{-P/2} exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)) \tag{129}$$

Posons

$$K = \Sigma^{-1} \tag{130}$$

La matrice de corrélation. On a

$$f_\alpha(x)\alpha|K|^{P/2} exp(-\frac{1}{2}(x-\mu)^T \Sigma(x-\mu)) \tag{131}$$

Définition

Un modèle graphique $G(V, E)$ où $V = 1, \ldots, P$ l'ensemble de noeudes et $E =$ l'ensemble des liens connectant certaine paire de noeud. Le paire $(i, j) \in E \Leftrightarrow X_i, X_j$ sont "conditionnellement dépendants " sachant toutes les autres variables $X_{V\{i,j\}}$ Autrement dit, (i, j) not $\in E \Leftrightarrow$ sont conditionnellement indépendantes sachant $X_{V\{i,j\}}$

Proposition

$$(i, j) \in E \Leftrightarrow K_{ij} \neq 0 \tag{132}$$

Preuve: Supposons $\mu = 0$, ainsi

$$f_x(x)\alpha exp(-\frac{1}{2}x^T Kx) \tag{133}$$

La densité conditionnelle de (X_i, X_j) sachant toutes autres variables est définie par

$$f(x_i, x_j|x_1, \ldots, x_p)\alpha exp(-\frac{1}{2}x_b^T K_{bb} x_b) \tag{134}$$

Avec

$$K_{bb} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \tag{135}$$

De dimension 2×2. Ainsi

$$f(x_b|x_*)\alpha exp(-\frac{1}{2}x_1^T K_{11}x_1 - \frac{1}{2}x_2^T K_{22}x_2 - x_1^T K_{12}x_2)a = f(x_1)f(x_2)exp(-x_1^T K_{12}x_2) \tag{136}$$

(Whatever)

Afin d'interpreter les élements de la matrice de ?, on étude maintenant la corrélation partielle.

Soit p_{ij} la corrélation entre X_i et X_j après avoir élinminé l'effet de toute les elements $\{X_k | k \in V$
$\{i, j\}\}$

2 Analyse de données structurées

$$X_1 \rightarrow y_1 = X_1 w_1, \ldots, X_J \rightarrow y_J = X_J w_J \tag{137}$$

$$\max_{w_1,\ldots,w_J} \sum_{j=1}^{J} \sum_{k=1}^{J} cos(X_j w_j, X_k w_k) \tag{138}$$

Rappel: Analyse en Composantes principales

Objectif: Trouver une combinaison linéaire des colonnes de X qui soit "représentative"

Critère: $w_1 = argmax_w var(Xw)$ s.t. $||w|| = 1$

Solution: w_1 est premier vecteur propre de $\frac{1}{n} X^T X$

$$\frac{1}{n} X^T X w = \lambda w \tag{139}$$

* Regression PLS-1

Objectif: Trouver une combinaison linéaire des colonnes de X: $t = Xw$ bien explicative de son propre bloc et corélé à y.

Critère:

$$w_1 = argmax_w cov(Xw, y) \text{ s.t. } ||w|| = 1 \tag{140}$$

Solution:

$$w_1 = \frac{X^T y}{||X^T y||} \tag{141}$$

Méthode d'analyse de données à 2 blocs L'objectif des méthodes d'analyse de données structurées en 2 blocs est de comprendre la relation. Il s'agit d'identifierdes sous-ensemble de variables dans chaque bloc qui "créent" le lien.

Une première méthode intitulée Régression PLS 2 est définie par le critère suivant

$y_1 = X_1 w_1$: y_1 composante, w_1 vecteur de poids $(w_1, w_2) = argmax_{w_1 \in \mathbb{R}^{P_1}, w_2 \in \mathbb{R}^{P_2}} cov(X_1 w_1, X_2 w_2)$ s.t. $||w_2|| = 2$

Pour résoudre ce problème d'optimisation, on passe par la fonction Lagrangien donnée par

$$L = \frac{1}{n} w_1^T X_1^T X_2 w_2 - \lambda_1 (w_1^T w_1 - 1) - \lambda_2 (w_2^T w_2 - 1) \tag{142}$$

On va dériver L par rapport à w_1 et w_2

$$\frac{\partial L}{\partial w_1} = \frac{1}{n} X_1^T X_2 w_2 - 2\lambda_1 w_1 \tag{143}$$

$$\frac{\partial L}{\partial w_2} = \frac{1}{n} w_1^T X_1^T X_2 - 2\lambda_2 w_2 \tag{144}$$

$$\begin{cases} \frac{1}{n} X_1^T X_2 w_2 = 2\lambda_1 w_1 \\ \frac{1}{n} X_2^T X_1 w_1 = 2\lambda_2 w_2 \end{cases} \tag{145}$$

En multipliant de par et d'autre du signe égal par w_1^T et w_2^T, on obtient

$$\begin{cases} \frac{1}{n} w_1^T X_1^T X_2 w_2 = 2\lambda_1 \\ \frac{1}{n} w_2^T X_2^T X_2 w_1 = 2\lambda_2 \end{cases} \tag{146}$$

Ainsi

$$\lambda_1 = \lambda_2 \tag{147}$$

En injectant une équation dans l'autre, on obtient que

$$\frac{1}{4n^2} X_1^T X_2 X_2^T w_1 = \lambda w_1 \tag{148}$$

$$\frac{1}{4n^2} X_2^T X_1 X_1^T X_2 w_2 = \lambda w_2 \tag{149}$$

Conclusion: w_1 est 1er vecteur propre de $X_1^T X_2 X_2^T X_1$, w_2 est 1er vecteur propre de $X_2^T X_1 X_1^T X_2$

Remarque: La régression PLS 2 s'appuie sur un critère de covariance

$$cov^2(X_1 a_1, X_2 a_2) = var(X_1 a_1) * cov^2(X_1 a_1, X_2, a_2) * var(X_2 a_2) \tag{150}$$

On cherche une composante $y_1 = X_1 a_1$ bien explicative de son propre bloc $\rightarrow ACP(x_1)$

On cherche une composante $y_2 = X_2 a_2$ bien explicative de X_2. Ainsi $ACP(X_2)$

Plutôt que de maximiser la covariance entre composantes on peut vouloir maximiser la corrélation (Hoteling, 1936) propose le critère suivant

$$\begin{aligned} (a_1, a_2) = argmax_{a_1,a_2} cor(X_1 a_1, X_2 a_2) = argmax \frac{cov(X_1 a_1, X_2 a_2)}{\sqrt{var(X_1 a_1)}\sqrt{var(X_2 a_2)}} \\ = argmax_{a_1,a_2} \frac{\frac{1}{n} a_1^T X_1^T X_2^T a_2}{\sqrt{\frac{1}{n} a_1^T X_1^T X_1 a_1}\sqrt{\frac{1}{n} a_2^T X_2^T X_2 a_2}} \end{aligned} \tag{151}$$

On remarque que la solution est invariante par changement d'échelle et donc on peut considérer le problème d'optimisation équivalent suivant

$$(a_1, a_2) = argmax_{a_1 \in \mathbb{R}^{P_1}, a_2 \in \mathbb{R}^{P_2}} cov(X_1 a_1, X_2 a_2) \tag{152}$$

s.t.

$$\begin{cases} var(X_1 a_1) = 1 \\ var(X_2 a_2) = 1 \end{cases} \tag{153}$$

Pour résoudre ce problème d'optimisation, on passe comme précédemment par le Lagrangien

$$L = \frac{1}{n}a_1^T X_1^T X_2 a_2 - \lambda_1(\frac{1}{n}a_1^T a_1^T X_1^T X_1 a_1 - 1) \tag{154}$$

En annulant les dérivées Lagrangien par rapport à a_1 et a_2

$$\begin{cases} \frac{1}{n}X_1^T X_2 a_2 = 2\lambda_1 \frac{1}{n} X_1^T X_1 a_1 \\ \frac{1}{n}X_2^T X_1 a_1 = 2\lambda_2 \frac{1}{n} X_2^T X_2 a_2 \end{cases} \tag{155}$$

En multipliant par a_1^T et a_2^T, il vient que $\lambda_1 = \lambda_2$

En injectant l'un dans l'autre

$$\frac{1}{4n^2}(X_1^T X_1)^{-1} X_1^T X_2 (X_2^T X_2)^{-1} X_2^T X_1 a_1 = \lambda^2 a_1 = \frac{1}{4n^2} Q_{12} a_1 \tag{156}$$

En conclusion: a_1 est $1er$ vecteur propre de Q_{12}, a_2 est 1er vecteur propre de Q_{21}.

Résumé générale: Jusqu'alors, on a vu 2 méthodes 2 blocs basés sur les unités suivants Critère 1:

$$cov(X_1 a_1, X_2 a_2) \text{ s.c. } ||a_1|| = ||a_2|| = 1 \tag{157}$$

Finalement, l'analyse canonique (CCA) et PLS2 sont basées sur la même fonction objective mais avec des contraints différents.

Dans la suite nous allons présenter un cadre unifiant les 2 méthodes, Pour ce faire, introduisons des paramètres $\tau_j \in [0,1], j = 1,2$ et considérons le problème d'optimisation suivant:

$$(a_1, a_2) = argmax_{a_1 \in \mathbb{R}^{P_1}, a_2 \in \mathbb{R}^{P2}} cov(X_1 a_1, X_2 a_2) \tag{158}$$

s.c.

$$\begin{cases} (1-\tau_1) var(X_1 w_1) + \tau_1 ||a_1||_2^2 = 1 \\ (1-\tau_2) var(X_2 w_2) + \tau_2 ||a_2||_2^2 = 1 \end{cases} \tag{159}$$

Ce problème d'optimisation définit l'analyse canonique régularisée.

Si $\tau_1 = \tau_2 = 0$, alors on retrouve l'analyse canonique. Si $\tau_1 = \tau_2 = 1$, alors pm retrrouve PLS2. Si $\tau_1, \tau_2 = 0$, alors le critère sous-jacent devient

$$\max_{a_1} var(X_1 a_1) cov^2(X_1 a_1, X_2 a_2) \tag{160}$$

s.c.

$$||a_1|| = 1, var(X_2 a_2) = 1 \tag{161}$$

L'analyse des redondance (Wollenberg, 1977) s'appuient sur ce dernier critère.

Regardons maintenant ce qu'il se oasse quand $\tau_1 \in (0,1)$ et $\tau_2 \in (0,1)$.

On peut montrer que a_1 et a_2 sont vecteurs propres des matrices suivantes

Reprenons le problème d'optimisation de CCA regularisée

$$\max cov(X_1 a_1, X_2 a_2) \tag{162}$$

	$\tau_1 = 0$	$\tau_1 = 1$
$\tau_2 = 0$	Analyse canonique	Analyse de redondance de X_1 sur X_2
$\tau_2 = 1$	Analyse des redondances de X_2 sur X_1	PLS2

s.c.

$$(1-\tau_j)var(X_ja_j)+\tau_j||a_j||_2^2 = 1, j = 1,2 = (1-\tau_j)\frac{1}{n}a_j^TX_j^TX_ja_j+\tau_ja_j^Ta_j = 1 = a_j^T[(1-\tau_j)\frac{1}{n}X_j^TX_j+\tau_ji_{P_j}]a_j \tag{163}$$

La solution de ce problème d'oiptimisation est obtenu en recherchant les vecteurs propres de (en utilisant les même recettes que précédamment)

$$((1-\tau_1)\frac{1}{n}X_1^TX_1+\tau_1I_{P_1})^{-1}X_1^TX_2((1-\tau_2)\frac{1}{n}X_2^TX_2+\tau_2I_{P_2})^{-1}X_2^TX_1 \tag{164}$$

$$\hat{\Sigma}_{11}X_1^TX_2\hat{\Sigma}_{22}X_2^TX_1 \tag{165}$$

On voit apparaître des estimations régularisées des Σ_{11} et Σ_{22}.

Par symmétrique, on a que a_2 est 1er vecteur propre de

$$\hat{\Sigma}_{11}X_1^TX_2\hat{\Sigma}_{22}X_2^TX_1 \tag{166}$$

Ainsi,

$$X_2^T[(1-\tau_2)\frac{1}{n}X_2X_2^T+\tau_2I_n]^{-1}X_1X_1^T[(1-\tau_1)\frac{1}{n}X_1X_1^T+\tau_1In]^{-1}X_2a_2 = \lambda a_2 \tag{167}$$

Remarque

On obtient deux formulations équivalentes pour obtenir a_1 et a_2. Une formulation primale qu'on utilisera quand $n > p_j$ et une formulation duale à utiliser si $n < p_j$.

En plus, en -pré-multipliant à gauche par X_2, on obtient le problème au valeur propre/vecteur propre suivant

$$X_2^T[(1-\tau_2)\frac{1}{n}X_2X_2^T+\tau_2I_n]^{-1}X_1X_1^T[(1-\tau_1)\frac{1}{n}X_1X_1^T+\tau_1In]^{-1}X_2a_2 = \lambda X_2a_2 \tag{168}$$

Et posons

$$K_j = X_jX_j^T \tag{169}$$

on obtient alors,

$$K_2[(1-\tau_2)\frac{1}{n}K2+\tau_2I_n]^{-1}K_1[(1-\tau_1)\frac{1}{n}K_1+\tau_1I_n]^{-1}y_2 = \lambda y_2 \tag{170}$$

On constate que pour calculer les composantes y_1 et y_2, il suffit de reconnaître uniquement les matrices de produits scalaires entre observations pour chaque bloc X_1 et X_2.

On étend de fait les méthodes cités précédemment au contexte des noyaux.

PLS $\rightarrow$ kernel PLS $\leftarrow$ Rosipal, 2001 CCA $\rightarrow$ kernel CCA RA $\rightarrow$ kernel Redundancy Analysis

Exemple illustratif Si y est uni-variée, lakernel PLS se réduit à

Dans ce cours, l'analyse de tableaux multiples est présenté au travers de l'analyse canonique généraliséerégularisée.(RGCCA) proposée en 2011(Tenenhaus & Tenenhaus, 2011).

RGCCA est défini par le problème d'optimisation suivant:

$$max_{a_1,\dots,a_S} \sum_{j=1}^{J}\sum_{k=1}^{J} c_{jk} g(cov(X_j a_j, X_k a_k)) \tag{171}$$

s.c.

$$(1-\tau_j)var(X_j a_j) + \tau_j ||a_j||_2^2 = 1, j = 1,\dots,J \tag{172}$$

où

$$c_{jk} = \begin{cases} 0 si X_j not \leftrightarrow \\ 1 si X_j \leftrightarrow \end{cases} \tag{173}$$

g: fonction convexe

On va maintenant étudier en détail ce problème d'optimisation qu'on peut re-exprimer comme:

$$\max_{a_1,\dots,a_J} f(a_1,\dots,a_J) = \sum_{j,k=1}^{J} a_{jk} g(\frac{1}{n} a_j^T X_j^T X_k a_k) \tag{174}$$

s.c.

$$(1-\tau_j)\frac{1}{n} a_j^T X_j^T X_j a_j + \tau_j a_j^T a_j = 1, j = 1,\dots,J = a_j^T[(1-\tau_j)\frac{1}{n} X_j^T X_j + \tau_j I_{p_j}]a_j = 1 = a_j^T M_j a_j = 1 \tag{175}$$

Posons $b_j = M_j^{1/2} a_j$ et $P_j = X_j M_j^{1/2}$

On obtient alors

$$\max_{b_1,\dots,b_J} f(b_1,\dots,b_J) = \sum_{j,k=1}^{J} c_{jk} g(\frac{1}{n} b_j^T P_j^T P_k b_k) \tag{176}$$

s.c.

$$b_j^T b_j = 1 \tag{177}$$

Pour résoudre ce problème, on va utiliser deux ingrédients 1. Block relatation 2. Majorization par minorization (MM)

Analyse de données multibloc

Ecrivons le Lagrangien assoué au problème d'optimisation de RGCCA

$$L = \sum_{j,k=1}^{J} c_{jk} g(\frac{1}{n} b_j^T P_j^T P_k b_k) - \sum_{j=1}^{J} \lambda_j (b_j^T b_j - 1) \tag{178}$$

Annulons la dérivée de L par rapport à b_j

$$\frac{\partial L}{\partial b_j}\sum_{k=1}^{J} c_{jk} g'(\frac{1}{n} b_j^T P_j^T P_k b_k) \frac{1}{n} P_j^T P_k b_k - 2\lambda_j b_j = 0 \tag{179}$$

$$b_j = \frac{1}{2\lambda_j}\sum_{k=1}^{J} c_{jk} g'(\frac{1}{n} b_j^T P_j^T P_k a_k) P_j^T P_k b_k \tag{180}$$

Posons

$$\epsilon_j = \sum_{k=1}^{J} c_{jk} g'(\frac{1}{n} b_j^T P_j^T P_k b_k) P_k b_k \tag{181}$$

Ainsi,

$$b_j = \frac{P_j^T \epsilon_j}{||P_j^T \epsilon_j||} \tag{182}$$

Or

$$b_j = M_j^{1/2} a_j \tag{183}$$

et

$$P_j = X_j M_j^{-1/2} \tag{184}$$

Ainsi,

$$a_j = \frac{M_j^{-1/2} M_j^{-1/2} X_j^T \epsilon_j}{\epsilon_j^T X_j^T M_j^1 X_j \epsilon_j} = \frac{M_j^{-1} X_j^T \epsilon_j}{\epsilon_j \dots} \tag{185}$$

Quelques commentaires

$$a_j \alpha [(1-\tau_j)\frac{1}{n} X_j^T X_j + \tau_j I_{P_j}]^{-1} X_j^T \epsilon_j \tag{186}$$

Si $\tau_j = 0$, ainsi $a_j \sim (X_j^T X_j)^{-1} X_j^T \epsilon_j \Leftrightarrow a_j$ est obtenu par régression multiple de ϵ_j sur X_j

Si $\tau_j = 1$, alors la contrainte devient $||a_j|| = 1$

$$a_j = \frac{X_j^T \epsilon_j}{||X_j^T \epsilon_j||} \tag{187}$$

....

Reprenons la forme générale pour a_j

$$a_j = [\epsilon_j^T X_j^T M_j^1 X_j \epsilon_j]^{-1/2} M_j^{-1} X_j^T \epsilon_j \tag{188}$$

$$a_j = [\epsilon_j^T X_j^T [(1-\tau_j)\frac{1}{n} X_j^T X_j + \tau_2 I_{P_j}]^{-1} X_j^T \epsilon_j]^{-1/2} [(1-\tau_j)\frac{1}{n} X_j^T X_j + \tau_j I_{P_j}]^{-1} \tag{189}$$

Sparse Partial Least Squares

$$\max cov(Xa, y) \tag{190}$$

s.c.

$$||a||_2 = 1 \tag{191}$$

$$||a||_1 < s \tag{192}$$

P1. Montrer que la solution optimale de SPLS est donnée par

$$a^* = \frac{S(\frac{1}{n}X^T y, \lambda_1)}{||S(\frac{1}{n}X^T y, \lambda_1)||_2} \tag{193}$$

où S est l'operation de seuillage doux.

Q2: Implémentation cet algorithme ou utiliser le package RGCCA(SGCCA).

Q3: Tester votre algorithme sur le jeu de données Alzhieimer

Q4: Par une procédure de déflation, construire une deuxième composante PLS.

Q5: Visualisation des individus sur le plan (y_1, y_2)

Q6: Comparer les résultats à ceux obtenus par les packages.