Introduction aux modèles graphiques **2011/2012**

Cours 6 — 2 novembre

Enseignant: Guillaume Obozinski *Scribe: François-Xavier Thomas, Martin Royer*

6.1 Max de vraisemblance *vs* max d'entropie

6.1.1 Dualité lagrangienne

On considère un problème $(\mathcal{P})$ d'optimisation de la forme suivante :

$$\min_{x\in\mathcal{X}} f(x) \quad \text{sous les contraintes} \quad Ax+b=0,$$

pour $\mathcal{X}$ un ensemble convexe et f une fonction convexe.

Pour se débarasser des contraintes, on utilise $H = \{x|Ax+b=0\}$ et on définit ι_H la function indicatrice de l'ensemble H :

$$\iota_H(x) = \begin{cases} 0 & \text{si} \quad Ax+b=0 \\ +\infty & \text{sinon.} \end{cases}$$

Résoudre le problème $(\mathcal{P})$ est donc équivalent à résoudre $(\mathcal{P}')$:

$$\min_{x\in\mathcal{X}} f(x) + \iota_H(x).$$

On définit le lagrangien $\mathcal{L}$ de ce problème défini comme une fonction de $x \in \mathcal{X}$ et d'un vecteur de variables duales $\lambda \in \mathbb{R}^n$, appelés aussi multiplicateurs de Lagrange :

$$\mathcal{L}(x,\lambda) = f(x) + \lambda^T (Ax+b)$$

On considère alors $\max_{\lambda\in\mathbb{R}^n} f(x) + \lambda^T(Ax+b)$. Si $Ax+b=0$, $\forall\lambda\in\mathbb{R}^n$, la quantité est nulle, et sinon on peut choisir λ avec des composantes du bon signe et tendant chacune vers $\pm\infty$ de telle sorte que

$$\max_{\lambda\in\mathbb{R}^n} f(x) + \lambda^T(Ax+b) = f(x) + \iota_H(x).$$

Le problème $(\mathcal{P})$ est donc résolu si on résout

$$\min_{x\in\mathcal{X}} \left[\max_{\lambda\in\mathbb{R}^n} \mathcal{L}(x,\lambda) \right].$$

L'inégalité suivante est toujours vraie et facile à vérifier :

$$\min_{x\in\mathcal{X}} \left[\max_{\lambda\in\mathbb{R}^n} \mathcal{L}(x,\lambda) \right] \geqslant \max_{\lambda\in\mathbb{R}^n} \left[\min_{x\in\mathcal{X}} \mathcal{L}(x,\lambda) \right]$$

On voudrait donc l'égalité. Cette inégalité devient une égalité dans certains cas, comme :

Théorème 6.1 (Théorème min-max pour les condition de Slater) *Si $\mathcal{X}$ convexe, et f convexe, et si $\exists x \in \text{Relint}(\mathcal{X})$ (intérieur relatif de $\mathcal{X}$ induit sur $\mathcal{X}$ par la topologie de l'enveloppe affine de X.), alors il y a égalité.*

De plus, les solutions du problème min-max sont les solutions du problème max-min.

6.1.2 Application au maximum d'entropie

On considère x une v.a. dans $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$ discret et une fonction vectorielle. $\varphi(x) \in \mathbb{R}^k$. Soit $\Delta = \{p \in \mathbb{R}^k \mid p(x_i) \geqslant 0, \sum_{x\in\mathcal{X}} p(x) = 1\}$ le simplexe correspondant à l'ensemble des distributions de probabilité sur $\mathcal{X}$ représentées comme le vecteur $p = (p(x_1), \ldots, p(x_m))^T$.

Etant donné un échantillon i.i.d. $(x^{(1)}, \ldots, x^{(n)})$, on souhaite trouver la distribution la plus "générale" satisfaisant la contrainte que l'espérance de ϕ pour p, $\mu_p = \mathbb{E}_p[\varphi(X)]$ coïncide avec sa valeur moyenne empirique observée sur les données. Une formalisation possible du problème est de chercher la distribution d'entropie maximum, donc en un sens ayant l'a priori le plus neutre possible sur les données, mais satisfaisant la contrainte souhaitée.

Cela conduit au problème :

$$\max_{p\in\Delta} H(p) \qquad \text{tel que} \qquad \mathbb{E}_p[\varphi(X)] = \bar{\varphi}$$

où on a noté $\bar{\varphi} = \frac{1}{n}\sum_{i=1}^n \varphi\left(x^{(i)}\right)$.

On veut donc maximiser :

$$\max_p \left[-\sum_x p(x)\log p(x)\right] \qquad \text{tel que} \qquad \sum_x p(x)\varphi(x) = \bar{\varphi}.$$

Le lagrangien associé à ce problème d'optimisation est

$$\mathcal{L}(p,(\lambda,c)) = -\sum_x p(x)\log p(x) + \lambda^T\left(\sum_{x\in\mathcal{X}} p(x)\varphi(x) - \bar{\varphi}\right) + c\left(\sum_x p(x) - 1\right),$$

où $(\lambda, c) \in \mathbb{R}^{k+1}$ sont les multiplicateurs de Lagrange.

Pour appliquer le théorème **??**, il faut vérifier la condition de Slater, i.e. vérifier qu'il existe $p \in \text{Relint}(\Delta)$ tel que $\mathbb{E}_p[\varphi(x)] = \bar{\varphi}$. Mais $p \in \text{Relint}(\Delta)$ si et seulement si $p(x_j) > 0$ pour tout j. Soit il existe un tel p, soit il existe un ensemble J tel que $(\mathbb{E}_p[\varphi(x)] = \bar{\varphi}) \Rightarrow (p(x_j) = 0,\ j \in J)$. Sans nuire à la généralité on peut donc considérer les distributions sur $\mathcal{X}' = \mathcal{X}\backslash\{x_j\}_{j\in J}$, i.e. dans le simplexe $\Delta' = \Delta \cap \{p \in \mathbb{R}^m \mid p(x_j) = 0, j \in J\}$ pour lesquelles par construction il existe toujours $p \in \text{Relint}(\Delta')$ satisfaisant les contraintes linéaires.

L'application du théorème permet d'intervertir min et max

$$\max_p \min_{c,\lambda} \mathcal{L}(x,(\lambda,c)) = \min_{c,\lambda} \max_p \mathcal{L}(x,(\lambda,c))$$

On maximise donc d'abord $\mathcal{L}$ en p. Comme on a

$$\frac{\partial \mathcal{L}}{\partial p(x)} = -\left[\ln p\left(x\right) + 1\right] + \lambda^T \varphi(x) + c \qquad (\star)$$

et que le maximum est atteint pour $\frac{\partial \mathcal{L}}{\partial p(x)} = 0$, on doit avoir

$$\ln p(x) = \lambda^T \varphi(x) - c' \qquad \text{avec} \qquad c' = 1 - c$$

ce qui implique $p(x) = \exp\left[\lambda^T \varphi(x) - c'\right]$. Sous les contraintes précédentes, l'optimisation impose $c' = A\left(\lambda\right)$, ce qui signifie qu'à l'optimum p^* appartient à la famille exponentielle de statistique exhaustive φ et de paramètre canonique λ.

Compte tenu de (**??**), on obtient alors que pour p^* en multipliant les deux membres de l'équation $\frac{\partial \mathcal{L}}{\partial p(x)}(p^*, \lambda) = 0$ par $p^*(x)$ et en sommant en x :

$$-\sum_{x\in\mathcal{X}} p^*(x) \ln p^*(x) + \sum_{x\in\mathcal{X}} \lambda^T \varphi(x) p^*(x) - A\left(\lambda\right) \sum_{x\in\mathcal{X}} p^*(x) = 0$$

Or, $\sum_{x\in\mathcal{X}} p^*(x) = 1$, donc :

$$-\sum_{x\in\mathcal{X}} p^*(x) \ln p^*(x) + \sum_{x\in\mathcal{X}} \lambda^T \varphi(x) p^*(x) - A\left(\lambda\right) = 0$$

En resubstituant dans l'expression du lagrangien on obtient donc (avec $p^*(x) = \exp\left[\lambda^T \varphi(x) - A\left(\lambda\right)\right]$),

$$\min_{\lambda,c} \max_{p} \mathcal{L}\left(p, \lambda, c\right) = \min_{\lambda} \left[-\lambda^T \bar{\varphi} + A\left(\lambda\right)\right] = \min_{\lambda\in\mathbb{R}^k} \left[-\frac{1}{n} \sum_{i=1}^{n} \log p\left(x^{(i)}\right)\right] \quad \text{t.q.} \quad p(x) = e^{\lambda^T \varphi(x) - A(\lambda)}.$$

On a donc montré le théorème suivant dans le cas où X est un variable aléatoire à valeur dans $\mathcal{X}$ discret.

Théorème 6.2 *Soit $(x^{(1)}, \ldots, x^{(n)})$ l'échantillon d'une variable aléatoire à valeurs dans $\mathcal{X}$ discret ou $\mathbb{R}^p$, l'estimateur du maximum d'entropie sous contrainte d'égalité de moments ($\mu_p = \bar{\varphi}$ pour une certaine fonction φ) est l'estimateur du maximum de vraisemblance dans la famille exponentielle de statistique suffisante $\varphi(X)$ (donc $p(x) = \exp\left[\lambda^T \varphi(x) - A\left(\eta\right)\right]$).*

Remarque. Le théorème se démontre plus généralement que pour $\mathcal{X}$ discret en utilisant le calcul variationel ou de l'analyse convexe. Notons que le théorème ne s'applique que pour une famille exponentielle canonique, pas sur des famille exponentielles courbes. On doit de plus avoir des données complètement observées.

Définition 6.3 *On dit qu'une famille exponentielle est régulière si son domaine Ω est ouvert.*

On considère $\mathcal{M} = \left\{\mu \in \mathbb{R}^k \mid \exists p, \mathbb{E}_p\left[\varphi(x)\right] = \mu\right\}$ où p est pris parmi l'ensemble des distribution de probabilités sur $\mathcal{X}$ (et pas seulement dans la famille exponentielle de statistique exhaustive φ).

Dans une famille exponentielle régulière, comme Ω est ouvert ∇A est définit en tout $\eta \in \Omega$ et on a :

$$\nabla A\left(\eta\right) : \left\{ \begin{array}{l} \Omega \to \mathcal{M} \\ \eta \to \mathbb{E}_\eta\left[\varphi(x)\right] \end{array} \right.$$

Que dire de cette fonction, est-elle injective ? Quelle est son image ?

Théorème 6.4 *Dans une famille exponentielle régulière, le gradient est injectif ssi la représentation est minimale.*

On montre d'abord que la minimalité est nécessaire. Par définition d'une famille non minimale $\exists \gamma, c$ tq $\gamma^T \varphi(x) + c = 0$.

Soient η_1 et $\eta_2 = \eta_1 + t\gamma$, avec t suff. petit pour que $\eta_2 \in \Omega$.

On a alors $p_{\eta_1} = \mathrm{e}^{-\eta_1^T \varphi(x) - A(\eta_1)}$ et

$$p_{\eta_2} = \mathrm{e}^{\eta_2^T \varphi(x) - A(\eta_2)} = \mathrm{e}^{\eta_1^T \varphi(X) - tc - A(\eta_2)} = \mathrm{e}^{\eta_1^T \varphi(X) - A(\eta_1)} = p_{\eta_1}.$$

D'où $\mu\left(\eta_1\right) = \mu\left(\eta_2\right)$.

Réciproquement, si φ minimale, alors nous avons prouvé dans le cours 5 que A est *strictement* convexe. Pour $\eta_1 \neq \eta_2$,

$$\begin{array}{rcll} A\left(\eta_1\right) & > & A\left(\eta_2\right) + \nabla A\left(\eta_2\right)^T \left(\eta_1 - \eta_2\right) & \text{par stricte convexité,} \\ A\left(\eta_2\right) & > & A\left(\eta_1\right) + \nabla A\left(\eta_1\right)^T \left(\eta_2 - \eta_1\right) & \textit{idem}\text{, d'où} \\ 0 & > & \left(\nabla A\left(\eta_2\right) - \nabla A\left(\eta_1\right)\right)^T \left(\eta_1 - \eta_2\right) & \text{en sommant les deux éq. préc., i.e. :} \\ 0 & > & \left(\mu\left(\eta_2\right) - \mu\left(\eta_1\right)\right)^T \left(\eta_1 - \eta_2\right) & \\ & \Rightarrow & \mu\left(\eta_1\right) \neq \mu\left(\eta_2\right) & \end{array}$$

D'où μ injectif. On montre également :

Théorème 6.5 *Pour une famille exponentielle régulière, le gradient est surjectif de $\Omega \to \mathring{\mathcal{M}}$, où $\mathring{\mathcal{M}}$ est l'intérieur de $\mathcal{M}$.*

6.2 Modèles Gaussiens

6.2.1 Forme canonique de la famille exponentielle

On considère une variable gaussienne dans $\mathbb{R}^p$: $X \sim \mathcal{N}(\mu, \Sigma)$ avec $\mu \in \mathbb{R}^p$, $\Sigma \in \mathbb{R}^{p\times p}$, $\Sigma \succ 0$. Sa densité s'écrit :

$$p(x, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$

On transforme cette fonction en une forme exp. canonique, en notant $\eta = \Sigma^{-1}\mu$ et $\Lambda = \Sigma^{-1}$.

$$\begin{aligned}
(x-\mu)^T \Sigma^{-1}(x-\mu) &= x^T\Sigma^{-1}x - x\mu^T\Sigma^{-1}x + \mu^T\Sigma^{-1}\mu \\
&= x^T\Lambda x - 2\eta^T x + \eta^T\Lambda^{-1}\eta \\
p(x, \mu, \Lambda) &= \exp\left[\eta^T x - \frac{1}{2}x^T\Lambda x - A(\eta, \Lambda)\right] \\
A(\eta, \Lambda) &= \frac{1}{2}\eta^T\Lambda^{-1}\eta + \frac{p}{2}\log 2\pi - \frac{1}{2}\log|\Lambda|
\end{aligned}$$

$\theta = \{\Lambda, \eta\}$ forme donc les paramètres canoniques, on a donc une statistique exhaustive de la forme :

$$\begin{pmatrix} x \\ -\frac{1}{2}xx^T \end{pmatrix} \to \begin{pmatrix} x \\ -\frac{1}{2}\operatorname{Vec}\left(xx^T\right) \end{pmatrix} = \varphi(x)$$

On note que la représentation n'est pas minimale. La matrice Λ est souvent appelée *matrice de précision.*

6.2.2 Espérance et covariance

On peut alors déterminer l'espérance et la covariance de X :

$$\begin{aligned}
\nabla_\theta A(\eta, \Lambda) &= \mathbb{E}_\theta[\varphi(X)] \\
&= \begin{pmatrix} \mathbb{E}_\theta[X] \\ -\frac{1}{2}\mathbb{E}_\theta\left[XX^T\right] \end{pmatrix} \\
\mathbb{E}_\theta[X] &= \nabla_\eta A(\eta, \Lambda) \\
&= \Lambda^{-1}\eta \\
&= \mu \\
-\frac{1}{2}\mathbb{E}_\theta\left[XX^T\right] &= \nabla_\Lambda A(\eta, \Lambda) \\
&= -\frac{1}{2}\Lambda^{-1}\eta\eta^T\Lambda^{-1} - \frac{1}{2}\Lambda^{-1} \\
&= -\frac{1}{2}\left[\mu\mu^T + \Lambda^{-1}\right]
\end{aligned}$$

On pouvait aussi calculer la covariance grâce à

$$\begin{aligned}\nabla^2_\theta A(\eta,\Lambda) &= \mathrm{Cov}\left[\varphi(X)\right] \\ &= \mathbb{E}_\theta\left[\varphi(X)\varphi(X)^T\right] - \mathbb{E}_\theta\left[\varphi(X)\right]\mathbb{E}_\theta\left[\varphi(X)\right]^T \\ &= \begin{pmatrix} \mathbb{E}_\theta\left[XX^T\right] & \cdots \\ \cdots & \cdots \end{pmatrix} - \mathbb{E}_\theta\left[\varphi(X)\right]\mathbb{E}_\theta\left[\varphi(X)\right]^T\end{aligned}$$

D'où,

$$\begin{aligned}\mathrm{Cov}\left[X\right] &= \mathbb{E}_\theta\left[XX^T\right] - \mu\mu^T \\ &= \nabla^2_\eta(\eta,\Lambda) \\ &= \Lambda^{-1} \\ &= \Sigma\end{aligned}$$

6.2.3 Loi marginale et loi conditionelle

On cherche à présent à décrire la distribution si on sépare la variable X en deux composantes. En particulier, on souhaite déterminer les lois marginales et conditionnelles de chacune des composantes. On note

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}, \quad \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}.$$

$$X_1 \sim\ ?, \quad X_2|X_1 \sim\ ?$$

Quelles gaussiennes ? Quels paramètres ?

$$p(x_1,x_2) = \frac{1}{(2\pi)^{p/2}\left|\Sigma\right|^{1/2}} \exp\left[-\frac{1}{2}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}^T \Lambda \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}\right]$$

Pour exprimer $p(x_1,x_2)$ comme $p(x_1)\,p(x_2|x_1)$, on a besoin d'expliciter Λ^{-1} en fonction de Σ^{-1}. On utilise le complément de Schur.

6.2.4 Complément de Schur

On considère la matrice par blocs $M = \begin{pmatrix} A & L \\ R & U \end{pmatrix}$. On peut commencer par la diagonaliser.

On multiplie M à droite et à gauche par deux matrices inversibles et triangulaires par blocs, D et G pour obtenir une matrice Δ :

$$\begin{aligned}\begin{pmatrix} I & 0 \\ -RA^{-1} & I \end{pmatrix} \times \begin{pmatrix} A & L \\ R & U \end{pmatrix} \times \begin{pmatrix} I & -A^{-1}L \\ 0 & I \end{pmatrix} &= D \times M \times G \\ &= \begin{pmatrix} I & 0 \\ -RA^{-1} & I \end{pmatrix} \times \begin{pmatrix} A & 0 \\ R & U - RA^{-1}L \end{pmatrix} \\ \Delta &= \begin{pmatrix} A & 0 \\ 0 & U - RA^{-1}L \end{pmatrix}\end{aligned}$$

Définition 6.6 *Le complément de Schur de A s'écrit $\left[M_{/A}\right] = U - RA^{-1}L$.*

Le complément de Schur de U est symétrique : $\left[M_{/U}\right] = A - LU^{-1}R$ (on aurait obtenu complément de Schur de U en multipliant à gauche par G et à droite par D).

Lemme 6.7 (Lemme du déterminant)

$$|\Delta| = |DMG| = |D|\,|M|\,|G| = |A| \times \left|\left[M_{/A}\right]\right| = |U| \times \left|\left[M_{/U}\right]\right|$$

Lemme 6.8 (Lemme de positivité) *Pour une matrice M symétrique, $M \succcurlyeq 0$ ssi $A \succcurlyeq 0$ et $\left[M_{/A}\right] \succcurlyeq 0$.*

(on peut remplacer les inégalités larges par des inégalités strictes)
En effet, $G = D^T$ $\forall x$, $x^T \Delta x \geqslant 0 \Leftrightarrow \forall x, \left(D^T x\right)^T M \left(D^T x\right) \geqslant 0$, d'où $\forall y, y^T M y \geqslant 0$.

Inversion de matrice par blocs et formule de Woodbury-Sherman-Morrison

Si M est inversible, alors $\Delta^{-1} = G^{-1}M^{-1}D^{-1}$, et $M = G\Delta^{-1}D$. Cela signifie que l'inverse de M peut s'écrire :

$$M^{-1} = \begin{pmatrix} I & -A^{-1}L \\ 0 & I \end{pmatrix} \times \begin{pmatrix} A^{-1} & 0 \\ 0 & \left[M_{/A}\right]^{-1} \end{pmatrix} \times \begin{pmatrix} I & 0 \\ -RA^{-1} & I \end{pmatrix}$$

En développant, on obtient la formule de Woodbury :

$$M^{-1} = \begin{pmatrix} A^{-1} + A^{-1}L\left[M_{/A}\right]^{-1}RA^{-1} & -A^{-1}L\left[M_{/A}\right]^{-1} \\ -\left[M_{/A}\right]^{-1}RA^{-1} & \left[M_{/A}\right]^{-1} \end{pmatrix}$$

En utilisant $\left[M_{/U}\right]$, on obtient de même :

$$M^{-1} = \begin{pmatrix} \left[M_{/U}\right]^{-1} & -U^{-1}R\left[M_{/U}\right]^{-1} \\ -\left[M_{/U}\right]^{-1}LU^{-1} & U^{-1} + U^{-1}R\left[M_{/U}\right]^{-1}LU^{-1} \end{pmatrix}$$

6.2.5 Loi marginale et loi conditionelle gausienne (suite)

On reprend alors le cas de la matrice de covariance de la loi gaussienne. En lui appliquant la formule de Woodbury, on obtient :

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}\left[\Sigma_{/\Sigma_{11}}\right]^{-1} \\ -\left[\Sigma_{/\Sigma_{11}}\right]^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1} \end{pmatrix}$$

On a donc $\Lambda_{22}^{-1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, et on peut donc réécrire la loi normale :

$$p(x_1,x_2) \propto \exp\left[-\frac{1}{2}\Big((x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)+(x_1-\mu_1)^T\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}(x_1-\mu_1)\ ...\right.$$
$$\left.-2\,(x_1-\mu_1)^T\Sigma_{11}^{-1}\Sigma_{12}\Lambda_{22}(x_2-\mu_2)+(x_2-\mu_2)^T\Lambda_{22}(x_2-\mu_2)\Big)\right]$$

On écrit alors $b = \Sigma_{21}\Sigma_{11}^{-1}(x_1-\mu_1)$, et on remplace. On obtient alors $p(x_1,x_2) = p(x_1)p(x_2|x_1)$ avec

$$p(x_1) = \frac{(2\pi)^{-p_1/2}}{|\Sigma_{11}|^{1/2}}\exp\left[-\frac{1}{2}\Big((x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)\Big)\right] \qquad \text{et}$$

$$p(x_2|x_1) = \frac{(2\pi)^{-p_2/2}}{\left|\left[\Sigma_{/\Sigma_{11}}\right]\right|^{1/2}}\exp\left[-\frac{1}{2}\Big((x_2-\mu_2-b)\left[\Sigma_{/\Sigma_{11}}\right]^{-1}(x_2-\mu_2-b)\Big)\right]$$

On a donc $X_1 \sim \mathcal{N}(\mu_1,\Sigma_{11})$, et $X_2|X_1 \sim \mathcal{N}\left(\mu_{2|1},\Sigma_{22|1}\right)$, en notant :

$$\begin{cases} \mu_{2|1} = \mu_2 + b = \mu_2 + \Sigma_{21}\times\Sigma_{11}^{-1}(x_1-\mu_1) \\ \Sigma_{22|1} = \Lambda_{22}^{-1} = \left[\Sigma_{/\Sigma_{11}}\right] \\ \Lambda_{22|1} = \Lambda_{22} \qquad \text{(le conditionnement est simple en paramétrisation canonique)} \\ \eta_{2|1} = \Lambda_{22|1}\times\mu_{2|1} = \Lambda_{22}\mu_2 - \Lambda_{21}(x_1-\mu_1) = \eta_2 - \Lambda_{21}x_1 \end{cases}$$

6.2.6 Zéros de la matrice de précision et propriété de Markov

On considère la loi conditionnelle $p(x_i,x_j|x_B)$, avec $B = \{1,\dots,p\}\setminus\{i,j\}$ et on note $I = \{i,j\}$.

On a alors :

$$\eta_I = \begin{pmatrix} \eta_i - \Lambda_{iB}x_B \\ \eta_j - \Lambda_{jB}x_B \end{pmatrix} \qquad \text{et} \qquad \Sigma_{II|B} = \Lambda_{II}^{-1} = \begin{pmatrix} \lambda_{ii} & \lambda_{ij} \\ \lambda_{ji} & \lambda_{jj} \end{pmatrix} = \frac{1}{|\Lambda_{II}|}\begin{pmatrix} \lambda_{jj} & -\lambda_{ji} \\ -\lambda_{ij} & \lambda_{ii} \end{pmatrix}$$

D'où $\mathrm{Cov}(X_i,X_j|X_B) = \dfrac{-\lambda_{i,j}}{|\Lambda_{II}|}$ ou encore $\mathrm{Corr}(X_i,X_j|X_B) = \dfrac{-\lambda_{ij}}{\sqrt{\lambda_{ii}\times\lambda_{jj}}}$. Donc si $\lambda_{ij} = 0$ alors $X_i \perp X_j \mid X_B$ et comme la distribution est gaussienne, cela implique que X_i et X_j sont indépendants sachant X_B

6.2.7 Lemme d'inversion de matrice

On revient sur une conséquence du lemme de Schur pour le calcul d'inverses que l'on rencontrera souvent en apprentissage (e.g. régression linéaire régularisée).

Lemme 6.9 *(Inversion de matrice) Soit* $X \in \mathbb{R}^{p\times n}$

$$\left(\mathrm{Id} + \lambda X^TX\right)^{-1} = \mathrm{Id} - \lambda X\left(\mathrm{Id} + \lambda XX^T\right)^{-1}X^T$$

On veut souvent inverser des matrices de type $\left(\mathrm{Id} + \lambda X^T X\right)^{-1}$, où X matrice de design à n lignes (échantillon iid) et p colonnes (paramètres) et typiquement $p \gg n$. C'est dans ce cas que le lemme d'inversion de matrice est utile car il permet de se ramener de l'inversion d'une matrice $p \times p$ à l'inversion d'une matrice $n \times n$.

Si on identifie $M = \begin{pmatrix} \mathrm{Id} & X \\ X^T & -\frac{1}{\lambda}\mathrm{Id} \end{pmatrix} = \begin{pmatrix} A & L \\ R & U \end{pmatrix}$, alors $\left[M_{/U}\right]^{-1} = \left(\mathrm{Id} + \lambda X^T X\right)^{-1}$ est l'expression que l'on souhaite calculer. En utilisant la formule de Woodbury pour le complément de Schur $\left[M_{/A}\right]$ on a l'identité

$$\left[M_{/U}\right]^{-1} = A^{-1} + A^{-1} L \left[M_{/A}\right]^{-1} R A^{-1}.$$

Dans notre cas, cela nous donne $\left[M_{/U}\right]^{-1} = \mathrm{Id} + X\left(-\frac{1}{\lambda}\mathrm{Id} - XX^T\right)^{-1} X^T$, d'où le résultat.

On a alors $\left[M_{/\mathrm{Id}}\right] = \left(\mathrm{Id} + \lambda X^T X\right)^{-1}$. L'avantage est qu'on cherche alors à inverser $\left[M_{/A}\right]$, une matrice $n \times n$ au lieu de $\left(\mathrm{Id} + \lambda X^T X\right)$, matrice $p \times p$.

On peut retrouver la formule en développant en série entière la formule initiale mais ça ne fait pas office de démonstration (valable seulement si la norme spectrale de X est strictement inférieure à 1, c'est-à-dire si la série converge).

6.3 Analyse en composantes principales

6.3.1 Formulation analytique

Données $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathbb{R}^{n \times p}$ où chaque ligne représente un échantillon iid, et les colonnes les descripteurs. On va chercher la direction de l'espace telle qu'une fois qu'on a projeté les données dans cette direction, la variance est maximale.

On veut donc $\max_u \mathrm{Var}_n\left(u^T x_i\right)$ (variance empirique des données) avec $\|u\|_2 = 1$.

Si on suppose les données centrées :

$$\frac{1}{n}\sum_{i=1}^{n} x_i = 0 \quad \text{d'où} \quad \mathrm{Var}\left(u^T x_i\right) = \sum_{i=1}^{n} u^T x_i x_i^T u = u^T X^T X u$$

Le maximum est donc atteint pour u vecteur propre associé à la plus grande valeur propre de $X^T X$. On peut itérer ce procédé pour trouver d'autres directions principales :

Déflation

On projette x_i sur l'orthogonal de u :

$$x_i' \leftarrow x_i - \left(x_i^T u\right) u, \quad \text{d'où} \quad X' \leftarrow X - Xvv^T = X\left(\mathrm{Id} - vv^T\right)$$

.

On obtient la séquence des *composantes principales* qui sont en fait les vecteurs propres de X^TX par ordre décroissant des valeurs propres associées.

Si on fait la décomposition en valeurs singulières (SVD) de X, en écrivant $X = USV^T$, avec U et V orthogonales et $S = \mathrm{Diag}(\sigma_1, \ldots, \sigma_n)$ avec $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_n$, on sait que $X^TX = VS^2V^T$, et les vecteurs de l'ACP cherchés sont donc les vecteurs trouvés par la SVD.

Composantes principales

Les composantes principales sont donc les k premières colonnes de V. Si on veut projeter les données sur cette base de composantes principales, on calcule $XV = USV^TV = US$.

Variables principales

Les variables principales sont les k premières colonnes de U, et une image de la projection des variables initiales sur les variables principales est donnée par $X^TU = VS$. (Notons que si les données sont centrées réduites, les vecteurs représentant les variables sont sur la sphère unité).

6.3.2 Formulation de synthèse

Un problème a priori différent conduit à une solution liée à l'ACP : étant donnée la matrice X, on voudrait trouver une matrice de rang faible $\tilde{X}$ qui approche bien les données. Soit en utilisant la norme de Froebenius :

$$\min_{\tilde{X}} \left\| X - \tilde{X} \right\|_F^2 \qquad \text{avec rang} \qquad \left(\tilde{X}\right) \leqslant k$$

La solution $\tilde{X}$ est obtenue en projetant X sur ses k premières composantes principales.

6.3.3 Modèles ACP probabiliste et ACP factorielle

Problème où les données sont dans un espace engendré par un certain nombre de vecteurs $\lambda_1, \ldots, \lambda_k$ rassemblés dans une matrice $\Lambda = \begin{pmatrix} \lambda_1 & \ldots & \lambda_k \end{pmatrix} \in \mathbb{R}^{p \times k}$.

Pour obtenir une formulation probabiliste on considère un vecteur gaussien $X \sim \mathcal{N}(0, \mathrm{Id})$ latent qui correspond aux données observables $Y \sim \mu + \Lambda X + \varepsilon$, avec $\varepsilon \sim \mathcal{N}(0, \Psi)$.

La loi de $Y|X$ s'écrit alors $Y|X \sim \mathcal{N}(\mu + \Lambda X, \Psi)$. La covariance jointe de (X, Y) est la matrice par bloc

$$\Sigma = \begin{pmatrix} \mathrm{Id} & \Lambda^T \\ \Lambda & \left(\Lambda\Lambda^T + \Psi\right) \end{pmatrix}$$

Comme on n'observe que Y, on note que Λ n'est identifiable qu'à un changement de base près.

Les modèles d'ACP probabiliste et factorielle estiment eux aussi $\tilde{X}$, mais via un algorithme EM. De plus l'analyse factorielle estime aussi la covariance du bruit Ψ, alors que l'ACP probabiliste fait l'hypothèse que cette covariance est isotrope, i.e. $\Psi = \sigma^2 I_d$.