Statistique Mathématique

Feuille 1 - 1ère partie: *Modèles statistiques*

Quelques éléments de correction pour éviter les pannes sèches

X.B.

N.B. Quand l'ordre des questions ne correspond pas à celui de l'énoncé, c'est que celui proposé ici a paru plus progressif en difficulté à l'auteur du corrigé. A tort ou à raison (vos remarques sont encouragées).

Exercice 1: ***Observations indépendantes mais non identiquement distribuées***

a) Le paramètre rassemble toutes les constantes numériques inconnues, donc:

$$\theta = \begin{pmatrix} \alpha \\ \beta \\ \sigma^2 \end{pmatrix} \in \Theta = \mathbb{R}^2 \times \mathbb{R}_+^*$$

Vraisemblance: l'indépendance des observations donne:

$$L(x\,;\theta) = \prod_i L(x_i\,;\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_i (x_i - \alpha - \beta\, z_i)^2 \right)$$

On notera au passage la possibilité d'une vision géométrique des choses:

$$\sum_i (x_i - \alpha - \beta\, z_i)^2 = \| x - \alpha \mathbf{1}_n - \beta\, z \|^2 \quad \text{où} \quad x = (x_i)_{i=1,\,n} \;,\; \mathbf{1}_n = (1)_{i=1,\,n} \;,\; z = (z_i)_{i=1,\,n} \in \mathbb{R}^n$$

b) α n'est plus le même pour toutes les observations. Mais les coefficients α_i ne peuvent être complètement libres pour autant, car il y aurait trop de paramètres inconnus. On contraint les α_i par:

$$\underset{(n,1)}{\alpha} = \underset{(n,s)}{B} \quad \underset{(s,1)}{a} \qquad \text{où } B = [b_1\,,\,\ldots\,,\,b_s] \text{ matrice connue}$$

Ici, a (donc α) sont inconnus. Le paramètre est donc:

$$\theta = \begin{pmatrix} a \\ \beta \\ \sigma^2 \end{pmatrix} \in \Theta = \mathbb{R}^s \times \mathbb{R} \times \mathbb{R}_+^*$$

$$L(x\,;\theta) = \prod_i L(x_i\,;\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_i (x_i - \alpha_i - \beta\, z_i)^2 \right)$$

La vision géométrique des choses s'avère très utile:

$$\sum_i (x_i - \alpha_i - \beta\, z_i)^2 = \| x - \alpha - \beta\, z \|^2 = \| x - B\,a - \beta\, z \|^2$$

c) Le rendement est aléatoire mais varie *en moyenne* linéairement en fonction de la dose d'engrais. Mathématiquement:

$$\mathrm{E}(X_i) = \alpha + \beta\, z_i$$

C'est un modèle linéaire de l'espérance de X conditionnellement à Z . Si $X_i \sim N(\alpha + \beta\, z_i\,,\, \sigma^2)$, on se retrouve dans la situation de la question (a).

d) La situation décrite ici est celle d'un plan d'expérience: les observations se distinguent par les valeurs de variables contrôlées (non aléatoires): ici, la variété du plant et la dose d'engrais. Le groupe j contient n_j

plants de la variété j. Le plant i de ce groupe reçoit la dose z_{ij}.

$$\forall i,j: \quad \mathrm{E}(X_{ij}) = a_j + \beta\, z_{ij}$$

Il faut voir les choses vectoriellement, comme très souvent:

$$\alpha = \begin{pmatrix} \begin{pmatrix} \alpha_{11} \\ \vdots \\ \alpha_{1,n_1} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \alpha_{J,1} \\ \vdots \\ \alpha_{J,n_J} \end{pmatrix} \end{pmatrix} = a_1 \begin{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \ldots + a_J \begin{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \\ \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = a_1 \mathbf{1}_1 + \ldots + a_J \mathbf{1}_J$$

où $\mathbf{1}_j$ représente la variable indicatrice du groupe j.

On note $B = [\,\mathbf{1}_1, \ldots, \mathbf{1}_J]$ et on se retrouve dans la situation de la question (b).

Exercice 2: ***Processus autorégressif: observations non indépendantes!***

Les choses se compliquent, puisque les X_i ne sont pas indépendants. La preuve, ils sont corrélés:

$$Cov(X_i, X_{i-1}) = Cov(\mu + \rho(X_{i-1} - \mu) + \varepsilon_i, X_{i-1}) = \rho\, Cov(X_{i-1}, X_{i-1}) = \rho\, V(X_{i-1}) \neq 0$$

Puisque les X_i ne sont pas indépendants, il est hors de question de faire le produit de leurs densités pour avoir celle de leur vecteur. On ne sait même pas si ce vecteur a une loi gaussienne (auquel cas il suffit de calculer son espérance et sa matrice de variance pour avoir la loi).

Donc il faut ruser. Le point de départ de la ruse est clair: ici, ce sont les ε_i qui sont indépendants et de même loi gaussienne. Leur vecteur ε suit donc une loi normale multidimensionnelle et très sympathique, calculatoirement parlant. Il faut donc se débrouiller pour exprimer le vecteur X en fonction de ε.

CA, c'est une *super*-indication. Vous le faites? Comment? Bon, bon ... Alors:

Déjà, nous n'allons pas trimballer μ partout, alors pour simplifier les écritures, on pose:

$$Y_i = X_i - \mu$$

Bon débarras. Dès lors, on voit que:

$$Y_0 = \varepsilon_0$$
$$Y_1 = \rho\, Y_0 + \varepsilon_1$$
$$\ldots\ldots\ldots\ldots$$
$$Y_n = \rho\, Y_{n-1} + \varepsilon_n$$

C'est déjà plus joli. D'autant plus qu'on montre ainsi par récurrence élémentaire que $\mathrm{E}(Y_i) = 0$. Mais le sommet de l'esthétique est comme souvent atteint par l'écriture matricielle, spécialiste de la synthèse des choses empilées (et/ou juxtaposées). En effet, en notant $Y = (Y_i)_{i=1,n}$, on a:

$$(1) \quad Y = A\, Y + \varepsilon \quad \text{, avec:}$$

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \rho & 0 & 0 & 0 & 0 \\ 0 & \rho & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \rho & 0 \end{pmatrix}$$

L'équation (1) s'écrit de façon équivalente:

$$(I_n - A)\, Y = \varepsilon \quad \Leftrightarrow \quad Y = (I_n - A)^{-1} \varepsilon \quad (2)$$

En effet, il suffit de développer $\det(I_n - A)$ par rapport à sa première ligne pour avoir, par une récurrence

immédiate:

$$\det(I_n - A) = 1$$

et donc l'inversibilité de $(I_n - A)$.

Deux voies s'offrent à nous pour continuer: la longue et sinueuse (ou chemin des pédofères), et le raccourci (ou chemin des ptérofères).

Voie longue et sinueuse:

L'équation 2, linéaire s'il en est, nous prouve que Y est un vecteur gaussien. Ouf, sauvés. Pour avoir sa densité, il suffit en effet de calculer son espérance et sa matrice de variance:

$$EY = (I_n - A)^{-1} E\,\varepsilon = 0$$

$$\Omega = V(Y) = (I_n - A)^{-1} V\,\varepsilon\ (I_n - A)^{-1\prime} = \sigma (I_n - A)^{-1}(I_n - A)^{-1\prime}$$

$$\Leftrightarrow \Omega^{-1} = \sigma^{-2} (I_n - A)'(I_n - A)$$

Il s'ensuit:

$$L(x;\theta) = \frac{1}{((2\pi)^n \det \Omega)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2} Y'(I_n - A)'(I_n - A) Y \right)$$

Oui, mais il est facile de démontrer que $\det(\Omega) = \sigma^{2n}$ (une bonne occasion de réviser les formules de base du déterminant). Et d'autre part, n'oublions pas que:

$$(I_n - A)\ Y = \varepsilon$$

Ce qui fait que finalement:

$$L(x;\varepsilon) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} \varepsilon'\varepsilon \right)$$

Ça alors. C'est exactement la densité du vecteur ε. C'est en général là qu'on se dit qu'il doit y avoir un raccourci, c'est pas possible.

Raccourci:

Comme $\det(I_n - A) = 1$ et que c'est le jacobien du changement de variable entre ε et Y, il est trivial que la densité de Y égale celle de ε. Et voilà. On se disait bien, aussi.

Rappel : pour un changement de variable bijectif $Y = g(X)$ où g est différentiable de matrice jacobienne $\Gamma(x)$, on a :

$$f^Y(y)|dy| = f^Y(g(x))|\Gamma(x)||dx| = f^X(x)|dx|$$