TP 3
Estimer l'intercorrélation entre variables aléatoires

Comment diminue la corrélation des pluies en fonction de la distance?

Partie 1

Comment identifier et mesurer des corrélations faibles

Une question-clé est souvent de décider si deux variables aléatoires sont ou non corrélées. Si la corrélation est "substantielle" (par ex. $r > 0.8$), la solution est simple et rapide. Par contre, si on s'attend à une corrélation faible (par ex. $r \sim 0.1$). la question est plus délicate et requiert une méthode appropriée. Ce problème se pose à chaque fois qu'on veut étudier un effet faible comme par exemple l'incidence du rayonnement électromagnétique des téléphones portables sur la santé (s'il y en a un). C'est le but de la première partie de montrer comment procéder dans de tels cas.

Pour une variable aléatoire comme par exemple la moyenne un intervalle de confiance I joue le même rôle qu'une barre d'erreur en physique. Il nous dit que pour un niveau de confiance de 95% si on répète l'expérience 100 fois, 95 résultats vont tomber dans I. Pour une corrélation, l'intervalle de confiance a bien sûr aussi cette même interprétation et c'est de cet aspect que s'occupe la question 1.

Mais pour une corrélation l'intervalle de confiance a de plus un sens supplémentaire. En effet, lorsqu'on calcule une corrélation entre deux variables X et Y, on souhaite en premier lieu répondre à la question: "Y a-t-il oui ou non une liaison entre X et Y?". Un critère souvent utilisé est le suivant:

> **Critère de corrélation** Si l'intervalle de confiance ne contient pas 0 on estime qu'il y a une corrélation significative entre les variables. Inversement, s'il contient 0 on estime qu'il n'y a pas de corrélation significative.

Dans la question 2 on s'occupe de ce second aspect de l'intervalle de confiance.

1 Vérification de l'intervalle de confiance

A et B sont deux variables aléatoires gaussiennes indépendantes de moyenne nulle et d'écart-type $\sigma = 1$. A partir d'elles on définit deux nouvelles variables par:

$$X = A + kB, \quad Y = A - kB \qquad 0 \leq k \leq 1$$

(a) Calculez analytiquement la corrélation r entre X et Y en fonction de k.

(b) L'intervalle de confiance (r_1, r_2) d'une corrélation r calculée avec un nombre de points n est

donné par les formules suivantes[1]:

$$r_1 = \tanh(Z - \alpha),\ r_2 = \tanh(Z + \alpha),\ Z = 0.5 \ln \frac{1+r}{1-r},\ \alpha = \frac{\beta(p)}{\sqrt{n-3}} \qquad (1)$$

$\beta(p)$ est un nombre qui dépend du niveau de confiance choisi. Par exemple (pour une distribution gaussienne):

$$p = 38\% :\ \beta = 0.50; \quad p = 80\% :\ \beta = 1.28; \quad p = 95\% :\ \beta = 1.96; \quad p = 99.9\% :\ \beta = 3.29$$

On veut s'assurer que ces formules sont bien correctes.
Pour cela, prenez $k = 1/3$ et simulez X et Y par tirage d'un nombre de valeurs de l'ordre d'une trentaine, calculez r par MATLAB et r_1, r_2 par la formule (1) en y injectant la valeur exacte de r, puis itérez 100 fois (ou 1000 fois), et faites compter par MATLAB le nombre de valeurs n_{12} comprises dans l'intervalle (r_1, r_2).
Pour un niveau de confiance de $p\%$ on s'attend à ce que $n_{12} \simeq p$. Faire le calcul pour $p = 38\%,\ 95\%, 99.9\%$ et résumer les résultats dans un tableau donnant en pourcentages les différences entre les valeurs observées et attendues.

2 Choix de l'intervalle de confiance

On note I la largeur de l'intervalle de confiance: $I = r_2 - r_1$. Dans l'application du critère de corrélation on peut se tromper de deux façons différentes:

- Soit estimer qu'il y a une corrélation (0 en dehors de I) alors qu'en fait il n'y en a pas; c'est ce qu'on appelle une *erreur de première espèce* ou encore un *faux positif*.
- Soit estimer qu'il n'y a pas de corrélation (0 dans I) alors qu'en fait il y en a une; c'est ce qu'on appelle une *erreur de seconde espèce* ou encore un *faux négatif*.

La formule (1) montre, comme il est d'ailleurs intuitivement évident, que lorsque $p \to 1$ la largeur I augmente en même temps que $\beta(p)$. Donc plus le niveau de confiance sera voisin de 100% plus I aura de chance de contenir 0 et donc plus le risque d'erreur de seconde espèce sera important.
Inversement, lorsque $p \to 0$, I devient plus étroit et on aura plus de chance de commettre une erreur de première espèce. On voit donc que le choix du niveau de confiance est une affaire à la fois importante et délicate. La question suivante illustre ce problème.

(a) Tirez deux variables indépendantes A et B chacune avec $n = 20$ valeurs. Choisissez d'abord $p = 99.9\%$ et après avoir calculé $\widehat{r}, r_1, r_2$ décidez s'il y a une corrélation au non. Faites une boucle itérant 100 fois ce test et faites comptabiliser par MATLAB le nombre de fois qu'il y a une corrélation significative.
Faites aussi calculer par MATLAB la moyenne r_m des 100 valeurs de $\widehat{r}$.
Faites de même avec $p = 38\%$.
Quelle conclusion en tirez vous?

(b) Simulez les deux variables X et Y avec $k = 0.9$ ce qui correspond a une corrélation d'environ 0.1, prenez pour chaque variable $n = 20$ valeurs, puis effectuez les mêmes opérations qu'à la question précédente.
Quelle conclusion en tirez-vous?

[1] Notez que Z est simplement la fonction réciproque de la fonction tangente hyperbolique. Cette fonction transforme l'intervalle $(-1, 1)$ en $(-\infty, \infty)$. Cela suggère la méthode utilisée pour trouver r_1, r_2. On a d'abord fait un changement de variable qui a transformé la corrélation en une variable définie sur $(-\infty, \infty)$; puis on a appliqué à cette nouvelle variable la méthode standard pour une variable gaussienne, après quoi on est revenu aux corrélations en appliquant la fonction $\tanh$.

(c) Fabriquez 10 séries X, Y de $n = 50$ nombres comportant 5 cas sans corrélation (c'est-à-dire A, B) et 5 cas avec une corrélation faible, c'est-à-dire X, Y avec $k = 0.9$.
Pour ces 10 cas trouvez le meilleur niveau de confiance c'est-à-dire celui qui donne le plus de réponses correctes lorsque vous itérez la procédure 10 fois.
Répétez la même recherche après avoir remplacé $n = 50$ par $n = 500$.
Pouvez-vous tirer une conclusion de ces deux tests?

(d) Pour finir, vérifiez que l'intervalle de confiance (RLO,RUP) donné par la commande *corrcoef* de MATLAB est identique à celui donné par la formule (1). Faites cette vérification pour $p = 38\%, 95\%, 99.9\%$.

3 Question-bonus facultative: Largeur de l'intervalle de confiance lorsque $r \to 1$ ou $n \to \infty$.

Lorsque $r \to 1$ ou $n \to \infty$ on voit de suite que $I \to 0$ mais on voudrait savoir si cette convergence vers 0 se fait rapidement ou lentement.

(a) r **tend vers** 1**.** Donnez à la corrélation entre X, Y des valeurs qui convergent vers 1 (par exemple de la forme $r = 1 - (1/3)^i$), construisez les variables X, Y avec les valeurs de k correspondantes, calculez r_1, r_2 et tracez I en fonction de r en échelle bilogarithmique. Quelle conclusion en tirez-vous?
Pouvez-vous retrouver ce résultat analytiquement par un développement limité au voisinage de $r = 1$?

(b) n **tend vers l'infini.** Donnez à n des valeurs croissantes et suivre la même procédure qu'à la question précédente.
Pouvez-vous également retrouver analytiquement le résultat obtenu graphiquement par un développement limité?

Partie 2

Liaison entre hauteurs de pluie en différent lieux

1 Lecture et visualisation des données.

(a) Faire lire a MATLAB le fichier des précipipations dans 9 villes (voir code à la fin). Ces chiffres donnent les hauteurs de pluie mensuelles exprimées en millimètres. Ainsi Berlin reçoit environ 50 mm de pluie par mois.
Pour pouvoir utiliser ces chiffres en parallèle on se limitera à l'intervalle 1865–1920. Afin de voir si les résultats obtenus dans la suite sont stables dans le temps on coupe cet intervalle de 56 ans en deux sous-intervalles $I_1,\ I_2$ de 28 ans chacun. Le point de départ de cette partie est donc constitués par deux ensembles de 6 vecteurs $P1_i$ et $P2_i$ chacun de dimension $12 \times 28 = 336$ mois.
(b) Pour voir l'allure de ces précipitations, représentez les chiffres de Berlin pour les 10 années 1865–1874. Afin d'avoir une vue plus claire Tracez des lignes verticales pointillées (par ex en rouge) pour séparer les années successives.

(c) Pour se faire une première idée des liaisons entre ces séries calculez la corrélation entre les deux premières villes ainsi que l'intervalle de confiance. Les séries sont-elles corrélées?

2 Réduction de l'écart-type par moyennage.

La moyenne de ces 9 séries aura un écart-type plus faible que l'écart-type moyen des séries individuelles mais, du fait qu'elles sont corrélées, la réduction sera plus faible que celle donnée par la loi en $1/\sqrt{n}$ qui n'est valable que si les séries sont non-corrélées. Rappelez la formule (F_1) vue en cours donnant l'écart-type σ_m de la moyenne pour des séries dont la corrélation moyenne est $\overline{r}$. Dans la

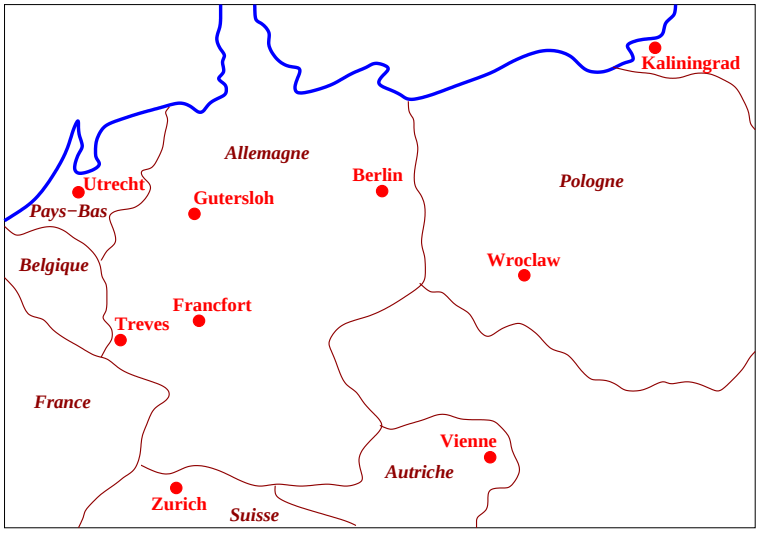

Fig. 1: Positions des villes pour l'étude des précipitations. Durant la période 1865–1920 Wroclaw et Kaliningrad faisaient partie de l'Allemagne sous les noms de Breslau et Königsberg.

question présente on veut voir si cette formule est en accord avec l'observation.

Dans tout ce qui suit on fera les calculs successivement pour I_1 et I_2 afin de pouvoir comparer les résultats.

(a) Avec les 9 villes on peut constituer $9 \times 8/2 = 36$ paires. Calculez les intercorrélations r_{ij} pour toutes les paires et obtenez ainsi la corrélation moyenne $\overline{r}$. En injectant cette valeur dans la formule (F_1), trouvez la valeur attendue de σ_m.

(b) Calculez la moyenne des 9 séries, puis son écart-type $\widehat{\sigma}_m$ et comparez à σ_m en exprimant la différence en %.

2 Calcul des distances entre villes

Nous avons déjà calculé les corrélations croisées r_{ij}; pour déterminer la fonction $r_{ij} = f(d_{ij})$ il ne reste donc qu'à obtenir les distances d_{ij}. Pour cela il nous faut d'abord introduire les latitudes et longitudes de chaque ville.

(a) Faire lire par MATLAB les fichiers des latitudes et longitudes puis convertissez ces chiffres en radians. Pour vérifier que ces chiffres sont corrects, représentez les 9 villes sur un graphique avec les longitudes en abscisses et les latitudes en ordonnées et comparez avec les positions de la Fig. 1.

(b) On note ϕ_i les longitudes et θ_i les compléments à $\pi/2$ des latitudes. A partir de ces angles et connaissant le rayon de la terre $R = 6,371$ km, on peut obtenir les distances $d_{ij} = M_iM_j$ entre les villes (i, j). Il suffit pour cela d'obtenir l'angle $\alpha = (\overrightarrow{OM_i}, \overrightarrow{OM_j})$ où O désigne le centre de la Terre; on aura: $M_iM_j = R\alpha$. Pour obtenir α il suffit de calculer le produit scalaire de deux vecteurs de longueur 1, $\overrightarrow{u_i}, \overrightarrow{u_j}$, portés respectivement par OM_i et OM_j. Pour calculer ce produit scalaire on va passer par les coordonnées de M_i et M_j. Celles-ci sont données par les formules suivantes (voir Fig. 2):

$$x = \sin\theta\cos\phi,\ y = \sin\theta\sin\phi,\ z = \cos\theta$$

A titre de vérification, calculez les distances Francfort–Trèves (150 km) et Berlin–Zurich (675 km)[2].

3 Évolution de l'intercorrélation avec la distance.

(a) Une fois les 36 distances et corrélations calculées, rajoutez à ces vecteurs un 37e point correspon-

[2]Comme dans les latitudes et longitudes on n'a tenu compte que des minutes et non des secondes la précision ne peut

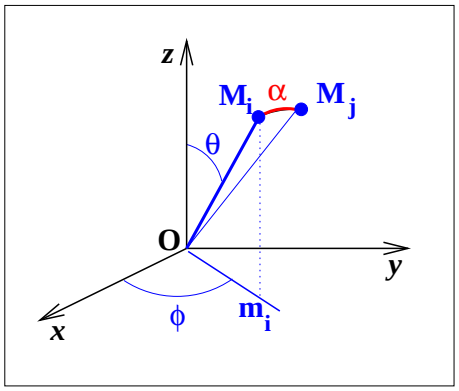

Fig. 2: Relation entre les angles θ et ϕ des coordonnées sphériques et les coordonnées x, y, z. La longueur Om_i est égale à: $Om_i = OM_i \sin(\theta)$; on en déduit de suite que:

$$x = OM_i \sin\theta \cos\phi,\ y = OM_i \sin\theta \sin\phi$$

dant à: $d = 0, r = 1$ (c'est-à-dire la corrélation d'une ville avec elle-même).
Puis représentez le nuage de ces 37 points sur un graphique où les distances sont en abscisses et les corrélations en ordonnées.

(b) On doit maintenant se demander quelle fonction il faut ajuster à ces points. La proposition la plus simple est une fonction linéaire décroissante $r_{ij} = -ad_{ij} + b,\ a > 0$. Mais une telle fonction ne peut pas convenir car pour des villes très éloignées on s'attend à ce que $r_{ij} \to 0$ mais non à ce que r_{ij} devienne négatif.
La fonction la plus simple satisfaisant à cette condition est: $r = \exp(-ad)$. Cette fonction a de plus l'avantage de donner la bonne valeur $r = 1$ pour $d = 0$.
Pour obtenir a il suffit de faire la régression linéaire entre $x = d$ et $y = \ln r$. Pour ne pas obtenir des valeurs de a trop petites, exprimez d en milliers de kilomètres. Est-ce que MATLAB donne un intervalle de confiance pour a? Calculez également la corrélation linéaire entre ces deux variables.

(c) En physique, plutôt que de garder la forme $r = \exp(-ad)$, il est d'usage d'écrire cette fonction sous la forme $r = \exp(d/\delta),\quad \delta = 1/a$. L'avantage de cette écriture est que δ a la dimension d'une longueur (alors que a a la dimension de l'inverse d'une longueur). Cette longueur est une caractéristique du système étudié. Lorsque la distance augmente de δ la corrélation est divisée par $e = 2.718$.
Quelles valeurs obtient-on ici pour δ.

4 Interprétation et généralisation

(a) Le calcul que nous venons de faire pour les précipitations en fonction de la distance peut être fait de la même façon pour bien d'autres variables. Citez-en quelques unes pour lesquelles on attend aussi une dépendance en fonction de la distance.

(b) A la fin du 19e siècle il se tenait un marché au blé chaque semaine (ou même plusieurs fois par semaine) dans pratiquement toutes les villes. Grâce à de telles statistiques on peut donc calculer un δ des prix du blé. A votre avis, ce δ sera-t-il plus grand ou plus petit que le δ des précipitations?

---

guère être meilleure que l'arc correspondant à un angle d'une demi-minute, (c'est-à-dire $1/120$ degré) soit:

$$R\left(\frac{1}{120}\right)\left(\frac{\pi}{180}\right) = 0.93 \text{ km}$$

Codes MATLAB

Partie 1

A vous de jouer!

Partie 2

```
%% Lecture des precipitations

% Ouverture du fichier
fIn = fopen('precp.data','r');
% Lecture de la premiere ligne
% et stockage dans la chaine de caractere tline
tline = fgetl(fIn);
iS=0;
iD=0;
% Initialisation d'un tableau de 1000 lignes et 13 colonnes:
% ce tableau est utilise pour le stockage temporaire des blocs de
% donnees (un bloc correspond aux precipitations pour une ville)
tmpD=zeros(1000,13);
while size(tline,2)>0
%Si le 1e caractere est un blanc, la ligne contient le nom d'une
  % ville
    if tline(1,1) == ' '
        % Mise en memoire du nom de la ville
        Name = '';
        iC=2;
        while tline(1,iC) ~= ' '
            Name = [Name tline(1,iC)];
            iC = iC +1;
        end
    else
        % Enregistrement d'une ligne de donnees
        iD = iD + 1;
        % Extraction des 13 valeurs et stockage dans le vecteur d
        d = sscanf(tline,'%f %f %f %f %f %f %f %f %f %f %f %f %f');
       % Copie de d dans le tableau
        tmpD(iD,:) = d;
    end
    % Lecture et stockage de la ligne suivante
    tline = fgetl(fIn);
  % Si ligne vide (fin des donnees) ou espace comme premier
    % caractere de la ligne (nouveau nom de ville) ...
    if (size(tline,2)==0 || tline(1,1)==' ')
   % On enregistre la serie de donnnees avec le nom de la ville
        iS = iS + 1;
      data(iS,1:2) = {Name tmpD(1:iD,:)};
      % puis recommence une nouvelle serie
```

```
        iD=0;
    end
end
fclose(fIn);

% Utilisation des donnees :
data{1,:} % Nom de la ville puis donnees
data{1,2} % Donnees seulement
```