Université Paul Sabatier
M2 SIA

1 Sujet 1 : description monodimensionnelle et bidimensionnelle de données

1.1 Première étude de la réflectance de matériaux

1.1.1 Données considérées

En utilisant la terminologie usuelle de la Statistique, on considère ici une population formée de n individus, pour lesquels on observe les valeurs numériques de p variables. Chaque individu est ici un bloc de matériau, d'un type spécifique. Chaque variable correspond à une longueur d'onde, et sa valeur pour un individu définit le degré avec lequel ce matériau réfléchit la lumière à cette longueur d'onde (la longueur d'onde augmente avec l'indice de la variable considérée). Au total, pour chaque individu, la suite de valeurs mesurées pour les variables considérées définit le spectre de réflectance de ce bloc de matériau.

Les données considérées sont disponibles dans le fichier igcp_1.sli . Elles peuvent être extraites de ce fichier à l'aide du programme Matlab load_igcp_1.m . On obtient ainsi une matrice X qui a la structure classique définie en cours.

Question : après avoir chargé les données, déterminer le nombre d'individus et le nombre de variables mis en jeu dans ces données.

1.1.2 Représentations partielles des données

A titre d'exemple, représenter la courbe donnant les valeurs de la variable d'indice $j = 1$, en fonction de l'indice i de l'individu considéré. Même question pour les variables d'indices 2, puis p.

A titre d'exemple, représenter le nuage de points bidimensionnel associé au sous-ensemble des données qui est formé seulement des valeurs prises par les variables d'indice 1 et 2 sur tous les individus. Même question en considérant les variables d'indices 1 et p.

Commenter les résultats, à la fois par rapport aux valeurs numériques des données considérées et par rapport aux propriétés physiques auxquelles on peut s'attendre dans le problème étudié.

A titre d'exemple, représenter la courbe donnant les valeurs des variables mesurées pour l'individu d'indice $i = 1$, en fonction de l'indice j de la variable considérée. Même question pour les individus d'indices 2, puis n. Commenter les résultats.

1.1.3 Analyse monodimensionnelle

1. Caractéristique de tendance centrale :
   Calculer la moyenne de chaque variable (sur tous les individus ; on leur affecte le même poids à tous dans tout ce T.P.). Pour cela, on réalisera deux programmes calculant chacun ces moyennes suivant une méthode différente, c.-à-d. :
   (a) à l'aide de l'expression matricielle de ces moyennes fournie en cours,
   (b) à l'aide de la fonction mean() de Matlab.

   Pour chacune de ces méthodes séparément, tracer les variations de la moyenne en fonction de l'indice de variable. Commenter les résultats en se référant à la nature des données considérées et aux résultats précédemment obtenus dans cette étude.

   Tracer la courbe donnant la différence entre les valeurs des moyennes calculées à l'aide de ces deux méthodes, et commenter la cohérence des résultats.

2. Caractéristique de dispersion :
   Calculer l'écart-type de chaque variable. Pour cela, on réalisera deux programmes calculant chacun ces écart-types suivant une méthode différente, c.-à-d. :
   (a) à l'aide de l'expression matricielle de ces écart-types fournie en cours,
   (b) à l'aide de la fonction std() de Matlab.

   Pour chacune de ces méthodes séparément, tracer les variations de l'écart-type en fonction de l'indice de variable. Commenter les résultats en se référant à la nature des données considérées et aux résultats précédemment obtenus dans cette étude.

   Tracer la courbe donnant la différence entre les valeurs des écart-types calculés à l'aide de ces deux méthodes, et commenter la cohérence des résultats.

1.1.4 Analyse bidimensionnelle

1. Calculer la matrice des coefficients de corrélation entre variables, en créant pour cela un programme personnel, c.-à-d. sans utiliser une fonction matlab qui donnerait directement le résultat.

   Tracer la courbe donnant les variations du coefficient de corrélation entre les variables d'indices 1 et j, en fonction de j.

   Ce résultat était-il prévisible, compte tenu des propriétés physiques auxquelles on peut s'attendre dans le problème étudié ?
2. On souhaite ici valider le résultat obtenu ci-dessus, en prenant en compte la totalité de la matrice de coefficients de corrélation disponible. Pour cela, on moyenne les coefficients de corrélation "adéquats". Plus précisément, au niveau élémentaire, on se donne une valeur k positive ou nulle de décalage entre les indices de deux variables, c.-à-d. on considère les variables d'indices j et $j + k$. A chaque valeur de j correspond ainsi la valeur du coefficient de corrélation des deux variables d'indices j et $j + k$. On calcule ensuite la valeur moyenne de ces coefficients de corrélation, pour toutes les valeurs de j qui sont acceptables pour la valeur de k considérée, c.-à-d. toutes les valeurs de j telles que les indices j et $j + k$ soient compris entre 1 et p. Ce coefficient de corrélation moyen dépend de k.

   On demande de calculer sa valeur pour chaque valeur possible de k (on définira dans le compte-rendu cette gamme de valeurs possibles), puis de tracer ses variations en fonction de k.

   Commenter les résultats.

1.2 Deuxième étude de la réflectance de matériaux

1.2.1 Données considérées

Les données considérées en Section 1.1 sont d'une nature particulière en ce sens que toutes les variables sont du même "type général" : ce sont des réflectances, en laissant ici de côté le fait que ces réflectances se rapportent à des longueurs d'onde différentes. Ceci permet d'étudier ici ces mêmes données dans une autre optique : on considère que les $n \times p$ valeurs contenues dans le fichier igcp_1.sli se rapportent à une unique variable (réflectance). On va ici étudier plusieurs propriétés statistiques de cette variable.

1.2.2 Histogramme

Réaliser un programme qui opère de la manière suivante :

1. Il détermine le domaine de variation de la variable, qui est borné par les valeurs minimale et maximale prises par cette variable sur l'ensemble des données considérées.
2. Il divise ce domaine en m intervalles, où m est un paramètre ajustable du programme. Ces intervalles ont la même largeur.
3. Il calcule les nombres de valeurs de la variable qui appartiennent à chacun de ces intervalles.

4. Il en déduit la fréquence de chacune des classes associées à ces intervalles.
5. Il déduit de ce qui précède l'histogramme de ces données pour $m = 10$. Même chose pour $m = 20$.

1.2.3 Moyenne et mode

Calculer la moyenne de la variable considérée.
Déterminer son mode pour $m = 10$, puis pour $m = 20$.
Comparer les résultats.

1.3 Etude de données synthétiques

1.3.1 Données considérées

Soient les données constituées des points suivants, qui correspondent chacun à un individu :
- tout d'abord, $2n + 1$ points définis comme suit. Chacun de ces points a pour indice i, avec $i = -n$ à n, et pour coordonnées (x_i^1, x_i^2). Ces coordonnées sont définies par :

$$x_i^j = \alpha_j \times i \quad \forall i \in \{-n, \ldots, n\}, \quad \forall j \in \{1, 2\} \tag{1}$$

  où α_1 et α_2 sont deux constantes réelles.
- en plus, un point ayant des coordonnées quelconques (y^1, y^2).

1.3.2 Coefficient de corrélation

1. Réaliser un programme qui opère de la façon suivante :
   (a) Pour des valeurs données de n, α_1, α_2, y^1, y^2, il génère les points de données définis ci-dessus.
   (b) Il représente graphiquement le nuage de points correspondant.

   On choisira les valeurs de n, α_1, α_2, y^1, y^2 de manière à ce que ce graphique soit facilement interprétable. On précisera dans le compte-rendu les valeurs ainsi choisies.
2. On considère des valeurs fixes de α_1, α_2, y^1, y^2, et on fait varier n. Réaliser un programme qui calcule explicitement les valeurs correspondantes du coefficient de corrélation des deux variables, à partir de toutes les coordonnées des points de données. Tracer l'évolution de ce coefficient de corrélation en fonction de n.

   A nouveau, on choisira les valeurs des paramètres de manière à ce que cette courbe soit facilement interprétable, et on précisera dans le compte-rendu les valeurs choisies.

   On considère maintenant l'expression littérale, établie en T.D, du coefficient de corrélation pour le type spécifique de données étudié ici. Tracer l'évolution de cette expression en fonction de n, dans les mêmes conditions que ci-dessus.

   Commenter tous les résultats.
3. On choisit $y^2 = 0$, on considère des valeurs fixes de n, α_1, α_2 et on fait varier y^1. Réaliser un programme qui calcule explicitement les valeurs correspondantes du coefficient de corrélation des deux variables. Tracer l'évolution de ce coefficient de corrélation en fonction de y^1.

   A nouveau, on choisira les valeurs des paramètres de manière à ce que cette courbe soit facilement interprétable, et on précisera dans le compte-rendu les valeurs choisies.

   Tracer aussi l'évolution de l'expression littérale du coefficient de corrélation, en fonction de y^1, dans les mêmes conditions que ci-dessus.

   Commenter tous les résultats.

2 Sujet 2 : analyse en composantes principales (ACP) et analyse factorielle discriminante (AFD)

2.1 ACP : première étude de la réflectance de matériaux

2.1.1 Données considérées

On considère à nouveau les données définies en Section 1.1.1. On complète ici leur étude à l'aide d'une ACP.

2.1.2 ACP avec métrique identité

On demande de réaliser un programme qui effectue l'ACP des données étudiées en utilisant la métrique $M = \mathrm{I}$ (sauf mention contraire, on affecte le même poids à tous les individus dans ce qui suit). Pour cela, on implantera les étapes de la méthode définie en cours (c.-à-d. on n'utilisera pas d'éventuelle fonction Matlab qui fournirait d'emblée tous les résultats de l'ACP). Dans ce programme, on pourra réaliser la diagonalisation de matrice à l'aide de la fonction eig() de Matlab. Ce programme devra fournir les résultats suivants :

1. Une courbe donnant chacune des valeurs propres associées aux axes principaux, par valeurs décroissantes, en fonction de l'indice de ces valeurs, avec une échelle logarithmique pour les ordonnées.

   Pour toutes les parties du programme mettant en jeu les valeurs propres, on utilisera en fait le *module* de ces valeurs car, même si ces valeurs sont réelles sur le principe pour l'étude réalisée ici, la méthode numérique employée dans la fonction Matlab eig() peut donner ici des valeurs de type complexe.
2. Une courbe donnant seulement les 10 premières valeurs propres associées aux axes principaux, par valeurs décroissantes, en fonction de l'indice de ces valeurs, avec une échelle linéaire pour les ordonnées.
3. Une courbe donnant le pourcentage d'inertie expliquée, en fonction de la dimension du sous-espace sur lequel on réalise la projection. On envisagera ici toutes les valeurs possibles de la dimension de ce sous-espace, par valeurs croissantes.
4. Une autre courbe du même type que ci-dessus, mais en ne considérant que les sous-espaces de dimension 1 à 10.

   Combien vaut ce pourcentage quand on projette les données sur un sous-espace de dimension 1 ? Même question pour un sous-espace de dimension 2, puis 3.

   Pour ces données, vous semble-t-il raisonnable de réaliser l'ACP en projetant les données sur un sous-espace de dimension 1 ? Même question pour un sous-espace de dimension 2, puis 3. Quelle est la motivation pour utiliser un sous-espace de dimension 2 ?
5. Une représentation du nuage projeté sur un sous-espace de dimension 1.
6. Une représentation du nuage projeté sur un sous-espace de dimension 2.
7. Les valeurs de la première composante principale.

Commenter les divers résultats ainsi obtenus, en indiquant les contraintes qui limitent éventuellement les possibilités d'interprétation de ces résultats.

2.2 ACP : deuxième étude de la réflectance de matériaux

2.2.1 Données considérées

Les données considérées ici ont la même structure générale qu'en Section 1.1.1, mais correspondent à trois classes différentes de matériaux, c.-à-d. à différents types de roches :

1. Un premier sous-ensemble de données est disponible dans le fichier ign_crs.sli . Elles peuvent être extraites de ce fichier à l'aide du programme Matlab load_ign_crs.m . Ces noms de fichiers proviennent du fait que ces données correspondent à des matériaux de type "Igneous Rocks - Coarse".

2. De même, les données disponibles dans le fichier ign_fn.sli peuvent être extraites à l'aide du programme load_ign_fn.m et correspondent à des matériaux de type "Igneous Rocks - Fine".
3. Enfin, les données disponibles dans le fichier sed_crs.sli peuvent être extraites à l'aide du programme load_sed_crs.m et correspondent à des matériaux de type "Sedimentary Rocks - Coarse".

Question : charger toutes ces données sous Matlab et les rassembler en une matrice X qui a la structure classique définie en cours. Déterminer le nombre d'individus et le nombre de variables mis en jeu dans chacune de ces classes de données et dans la base de données complète.

2.2.2 Représentations partielles des données

A titre d'exemple, représenter le nuage de points bidimensionnel associé au sous-ensemble des données qui est formé seulement des valeurs prises par les variables d'indice 1 et 2 sur tous les individus. On représentera les individus associés aux différentes classes par des pictogrammes différents.

Même question en considérant les variables d'indices 1 et p.

Commenter les résultats.

2.2.3 ACP avec métrique identité

Réaliser l'ACP de la base complète de données associée à la matrice X, suivant le même principe qu'en Section 2.1.2, à part pour les aspects suivants :

1. Pour les nuages projetés : ici, on représentera les individus associés aux différentes classes par des pictogrammes différents.
2. On ne donnera pas ici les valeurs de la première composante principale.

2.3 AFD : étude de la réflectance de matériaux

2.3.1 Données considérées

On considère à nouveau les données définies en Section 2.2.1. On complète ici leur étude à l'aide d'une AFD.

2.3.2 AFD

On demande de réaliser un programme qui calcule la version centrée Y des données initiales X, puis qui réalise l'AFD de ces données Y. Pour cela, on implantera les étapes de la méthode définie en cours (c.-à-d. on n'utilisera pas d'éventuelle fonction Matlab qui fournirait d'emblée tous les résultats de l'AFD). Dans ce programme, on pourra réaliser la diagonalisation de matrice à l'aide de la fonction eig() de Matlab.
Ce programme devra fournir les résultats suivants :

1. Une courbe donnant chacune des valeurs propres associées aux axes (et facteurs) discriminants, par valeurs décroissantes, en fonction de l'indice de ces valeurs, avec une échelle logarithmique pour les ordonnées.

   A nouveau, pour toutes les parties du programme mettant en jeu les valeurs propres, on utilisera en fait le *module* de ces valeurs.
2. Une courbe donnant seulement les 10 premières valeurs propres associées aux axes (et facteurs) discriminants, par valeurs décroissantes, en fonction de l'indice de ces valeurs, avec une échelle linéaire pour les ordonnées.
3. Une représentation du nuage non centré X projeté sur un sous-espace de dimension 1, en utilisant un pictogramme différent pour chaque classe.
4. Une représentation du nuage non centré X projeté sur un sous-espace de dimension 2, en utilisant un pictogramme différent pour chaque classe.

Commenter les divers résultats ainsi obtenus.

TP3 : Tests statistiques

Il est indispensable d'avoir préparé ce TP avant la séance. Tous les calculs théoriques, et en particulier le calcul des seuils utilisés dans les tests, doivent être faits avant la séance de TP.

I. Variables aléatoires

Les fonctions Matlab *rand* et *randn* permettent de générer deux variables aléatoires (V.A.) respectivement uniformément distribuée sur l'intervalle [0,1] et gaussienne centrée réduite (de moyenne nulle et de variance unité).

1. Écrire un programme Matlab permettant de générer 1000 réalisations d'une variable aléatoire uniformément répartie dans l'intervalle [-1 , 2]. Estimer la moyenne et la variance de cette variable en utilisant les fonctions Matlab *mean* et *var*. Comparer avec des valeurs théoriques.
2. Générer 1000 réalisations d'une V.A. gaussienne de moyenne 5 et de variance 3. Estimer la moyenne et la variance sur les réalisations générées et comparer avec les valeurs théoriques. Répéter l'expérience avec 10 réalisations seulement. Conclusion ?

II. Théorème central-limite

1. Générer une matrice 30*1000 des réalisations d'une variable distribuée sur l'intervalle [0,1]. Tracer l'histogramme de la première ligne de cette matrice en utilisant la fonction *hist* de Matlab.
2. Calculer la somme sur les colonnes de cette matrice et tracer l'histogramme du résultat. Conclusion ?

III. Tests sur les variances et moyennes d'échantillons indépendants

Les fichiers *y.mat* et *z.mat* contiennent chacun 40 mesures empiriques de deux variables aléatoires indépendantes.

1. Tester si la variance de chaque variable est égale à 0.4 ou non (test bilatéral de variance avec un niveau de signification de α=0.05).
2. En supposant que la variance de chaque variable est égale à 0.4, tester l'égalité de leurs moyennes (test bilatéral de Z avec α=0.05). Quelle est la valeur P pour ce test ?
3. Refaire le test de la question 2 en supposant la variance inconnue (test t de Student).

Remarque : Pour calculer la moyenne et la variance empiriques, on peut utiliser les fonctions Matlab *mean* et *var*.

IV. Comparaison des moyennes d'échantillons appariés

On souhaite vérifier si la concentration lipidique se modifie en conservant le sang pendant un certain temps. Les échantillons de sang de 20 sujets d'une certaine population ont été analysés immédiatement après la prise de sang et 8 mois après. Les résultats ont été stockés dans les deux lignes d'une matrice, enregistrée dans le fichier *sang.mat*. Utiliser un test de Student avec un niveau de signification de 0.05.

V. Test de corrélation

Le fichier *x.mat* contient N=31 échantillons temporels d'un bruit blanc gaussien x(t). On veut vérifier si deux échantillons successifs de ce bruit sont corrélés on non. On procède de la manière suivante.

1) Créer deux vecteurs x1 et x2, le premier contenant les échantillons d'indices 1 à 30 et le deuxième contenant ceux d'indices 2 à 31 du signal x(t).
2) Calculer le coefficient de corrélation empirique entre ces deux vecteurs en utilisant la fonction Matlab *corrcoef*.
3) Utiliser un test de Student avec un niveau de signification de 5% pour vérifier si les deux vecteurs sont corrélés ou non.
4) Filtrer le signal x(t) par le filtre RII passe-bas ci-dessous (en utilisant la fonction Matlab *filter*) pour obtenir un bruit coloré y(t). Refaire le test de corrélation sur ce nouveau signal.

$$y[n] = x[n] + 0.4 * y[n-1]$$

5) Justifier le résultat obtenu.

VI. Tests d'adéquation

On veut vérifier si la variable générée par la fonction *rand* de Matlab suit vraiment une loi uniforme sur [0 1].

1. Créer un vecteur de 60 réalisations aléatoires en utilisant la fonction *rand* de Matlab[1].
2. En utilisant un test du chi-2 avec 6 classes et avec un niveau de signification de 5%, tester l'adéquation des données avec une loi uniforme.
3. Répéter en utilisant un test de Kolmogorov-Smirnov.
4. Un étudiant distrait se trompe lors de la génération des réalisations de la variable ci-dessus et écrit dans son programme *randn* au lieu de *rand*. Exécuter votre programme dans ce cas et vérifier les résultats des tests du chi-2 et de Kolmogorov-Smirnov.

VII. Test du rapport de vraisemblance

Le fichier *Rayleigh.mat* contient 100 réalisations d'une variable suivant la loi de Rayleigh ci-dessous :

$$f_X(x) = \frac{2x}{b}\exp(\frac{-x^2}{b}) \quad , \quad x \geq 0 \quad , \quad b \geq 0$$

On veut décider entre les deux hypothèses suivantes :
H0 : b=2 H1 : b=1.5

En utilisant le test du rapport de vraisemblance, prendre une décision avec un risque de 1ère espèce α=0.05.

Remarque : Si X est une variable aléatoire suivant une loi de Rayleigh de paramètre b, la moyenne et la variance de la variable X^2 sont respectivement égales à b et b^2.

[1] Ajouter la commande *rand('seed',1)* avant de générer les réalisations de la variable.