Université de Lorraine - Site de Metz - UFR MIM **Année 2012/2013**
Master de Mathématiques, année M1
Module *Introduction à l'analyse des données*

Exercices - Feuille 1

Représentation graphique des données

1- Représentation graphique de données multivariées (1)
On souhaite représenter à l'aide de différentes méthodes graphiques les tableaux de données X correspondant aux trois jeux de données `pullover.dat`, `mandible.dat`, `frenchfood.dat`. Pour le tableau X provenant de `mandible.dat`, on obtient par exemple les figures Fig.1, Fig.2, Fig.3.

Donner quelques conclusions qualitatives déduites de ces représentations, concernant la corrélation entre les différentes variables observées.

Pour chacun des autres jeux de données, effectuer le même travail. On pourra utiliser les logiciels `R` ou `matlab`. En `matlab`, les commandes utiles sont `andrewsplot`, `glyphplot`, `boxplot`, `gplotmatrix`, `gscatter`.

2- Représentation graphique de données multivariées: billets de banque
On considère les données `swiss.dat` (six dimensions de billets de banque suisses.) Les 100 premiers billets sont authentiques et les 100 suivants sont des billets contrefaits (fausse monnaie). On veut savoir si il est possible de distinguer visuellement l'échantillon des billets authentiques de celui des billets contrefaits.
1) Lire en matlab le fichier des données. (Utliser `p7.m`).
2) Verifier les caractéristiques du billet numéro 86. On utilisera `bnote(86,:)`, `length(86)`, etc...
3) Tracer le scatterplot de chacun des échantillons de billets numérotés 1 à 50, puis 51 à 100, puis 101 à 151, enfin 151 à 200.
4) Tracer les visages de Chernoff de chacun des quatre échantillons, cf Fig.4, 5, 6, 7.
5) Tracer les diagrammes en étoile des deux échantillons.
6) Tracer les courbes d'Andrew correspondant aux données des billets 96 à 105. On tracera d'abord les données dans l'ordre 1 à 6, puis ensuite avec l'ordre de 6 à 1, cf Fig. 8.
7) Peut-on déduire un renseignement qualitatif de ces représentations ?

3- Données de Bumpus: survie de moineaux
1) Lire dans `matlab` les données de Bumpus `bumpus.dat` relatives à la survie de 49 moineaux après une tempête. On distinguera les données relatives aux 21 premiers oiseaux qui ont survécu et des oiseaux 22 à 49 qui n'ont pas survécu.
2) Représenter à l'aide des visages de Chernoff et de la représentation en étoile chacun des deux jeux de données.
3) Calculer la matrice de covariance de l'ensemble des données, puis de chacun des deux jeux de données.
4) Calculer aussi la moyenne et l'écart-type de chacun des deux jeux de données.

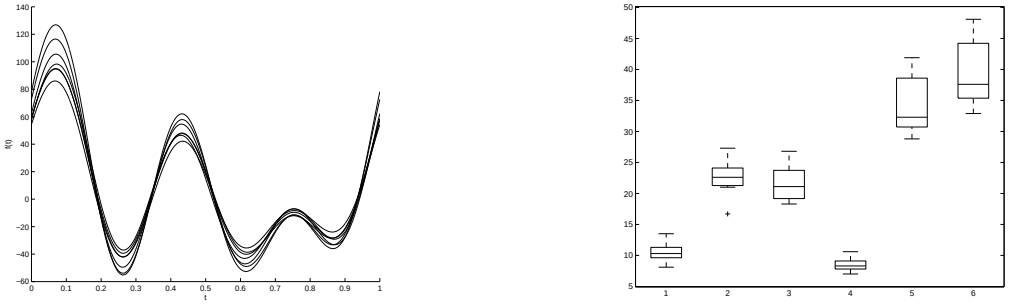

Figure 1: Courbes d'Andrew, Boxplots, `mandible.dat`

Figure 2: Représentation en étoiles, visages de Chernoff, `mandible.dat`

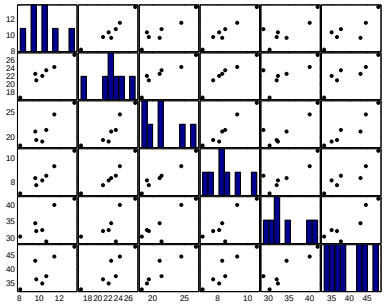

Figure 3: Scatterplots, `mandible.dat`

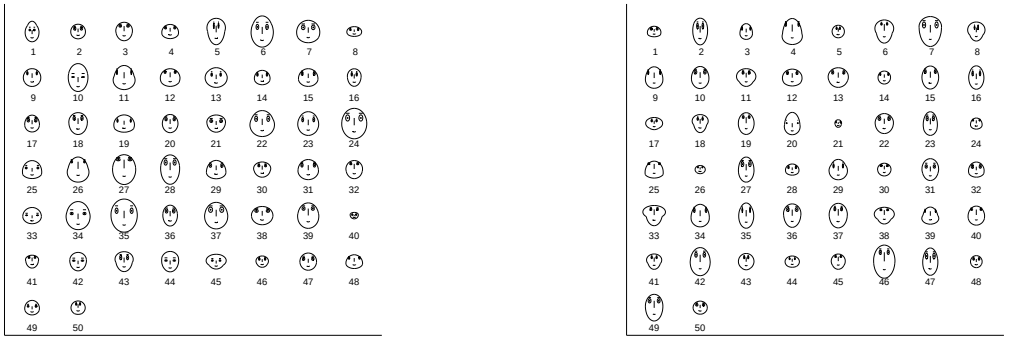

Figure 4: Billets de banque 1 à 50 (gauche) et 51 à 100 (droite) : Visages de Chernoff, `swiss.dat`

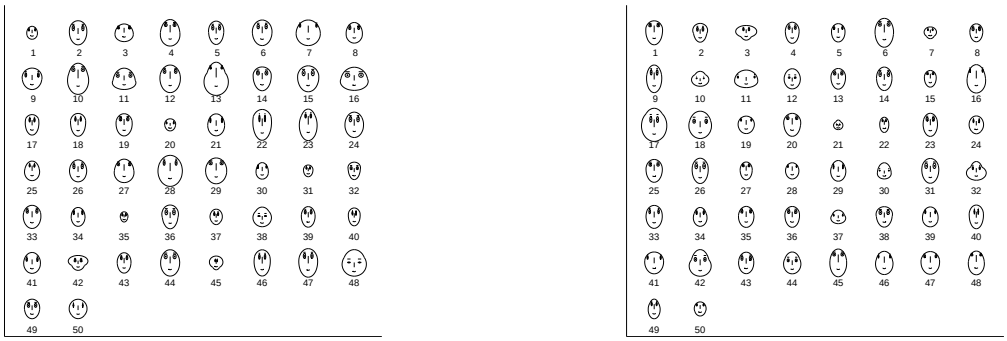

Figure 5: Billets de banque 101 à 150 (gauche) et 151 à 200 (droite) : Visages de Chernoff, `swiss.dat`

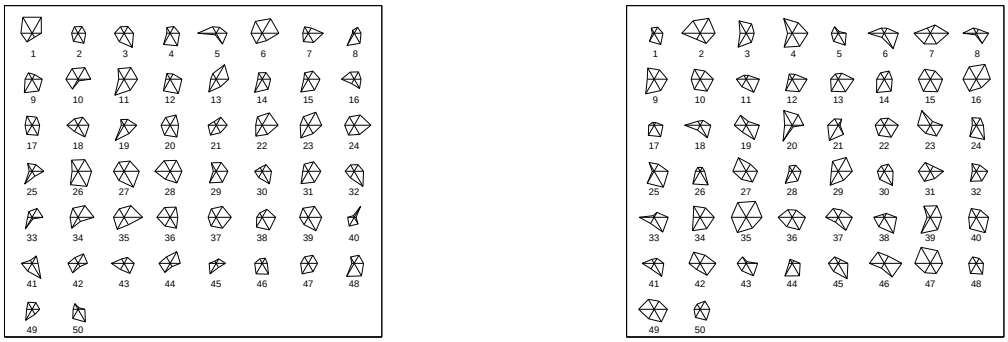

Figure 6: Billets de banque 1 à 50 (gauche) et 51 à 100 (droite) : Représentation en étoiles, `swiss.dat`

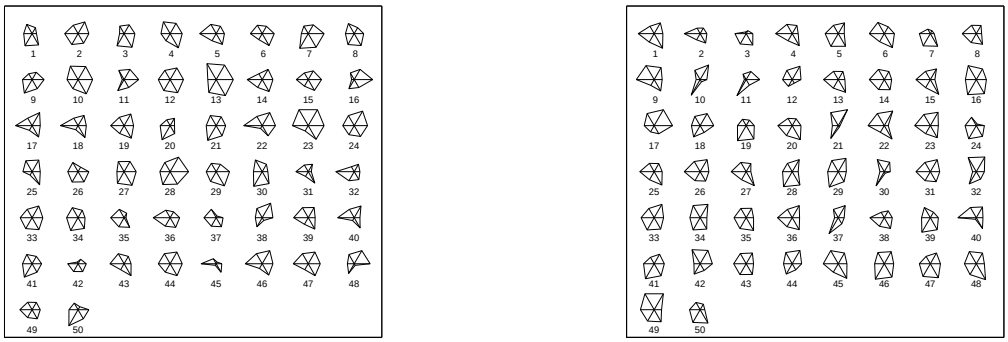

Figure 7: Billets de banque 101 à 150 (gauche) et 151 à 200 (droite) : Représentation en étoiles, `swiss.dat`

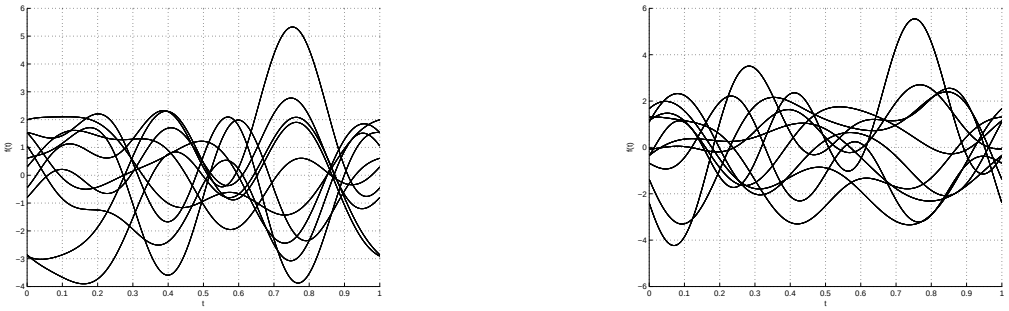

Figure 8: Billets de banque 96 à 105, données de 1 à 6 (gauche), et données de 6 à 1 (droite) : Courbes d'Andrew.

4- Vente de pull-overs bleus

Härdle et Simar: Applied Multivariate Statistical Analysis, Springer 2003.

Il s'agit d'un jeu de de 10 mesures avec 4 caractéritiques. Il s'agit du responsable d'un commerce de textiles étudie la vente de pulls bleus sur 10 périodes distinctes. Il utilise trois méthodes de marketing différentes. Il veut interpréter le résultat des ventes en fonction des données dont il dispose.

période	**prix**	**ventes**	**publicité**	**temps (en h.)**
1	230	125	200	109
2	181	99	55	107
3	165	97	105	98
4	150	115	85	71
5	97	120	0	82
6	192	100	150	103
7	181	80	85	111
8	189	90	120	93
9	172	95	110	86
10	170	125	130	78

1) Lire les données en `matlab`.
2) Calculer la matrice de covariance.
3) Calculer la moyenne et l'écart-type des données.

5- Chiens préhistoriques en Thaïlande

B.F.J. Manly, Multivariate Statistical Methods, a primer, Chapman& Hall, 2005.

On veut comprendre les corrélations entre espèces vivantes ou disparues et ainsi améliorer la construction des arbres phylogénétiques (genéalogie des espèces au cours du temps). Un exmple important est l'origine chien moderne. Des fouilles de sites préhistoriques en Thaïlande ont permis d'extraitre des os de canidés (chiens, loups,...). Une étude a donné le tableau suivant à partir duquel on veut effectuer des corrélations avec différents canidés:

- le chien moderne
- le chacal doré
- le loup chinois
- le loup indien

- le cuon
- dingo

Les données anatomiques sont les suivantes

- X_1 = largeur mandibule
- X_2 = hauteur mandibule
- X_3 = longueur de la première molaire
- X_4 = largeur de la première molaire
- X_5 = distance de la première molaire à la troisième molaire
- X_6 = distance de la première molaire à la quatrième molaire

	X_1	X_2	X_3	X_4	X_5	X_6
ch.moderne	10	21	19	8	32	37
chac.doré	8	17	18	7	30	33
l. chinois	14	27	27	11	42	48
l. indien	12	24	25	9	40	45
cuon	11	24	21	9	29	38
dingo	10	23	21	8	34	43
ch.préhist.	10	22	19	8	32	35

1) Lire les données en `matlab`.
2) Calculer la matrice de covariance.
3) Calculer la moyenne et l'écart-type des données.

6- Alimentation en France
Härdle et Simar: Applied Multivariate Statistical Analysis, Springer 2003.
Les données décrivent les valeurs moyennes de consommation alimentaire en France de 7 types d'aliments par 12 types de familles. Les abbréviations sont les suivantes: MA = ouvriers, EM = employés, CA = cadres. Le chiffre donne le nombre d'enfants dans la famille.

	pain	**légumes**	**fruits**	**viande**	**volaille**	**lait**	**vin**
MA2	332	428	354	1437	526	247	427
EM2	293	559	388	1527	567	239	258
CA2	372	767	562	1948	927	235	433
MA3	406	563	341	1507	544	324	407
EM3	386	608	396	1501	558	319	363
CA3	438	843	689	2345	1148	243	341
MA4	534	660	367	1620	638	414	407
EM4	460	699	484	1856	762	400	416
CA4	385	789	621	2366	1149	304	282
MA5	655	776	423	1848	759	495	486
EM5	584	995	548	2056	893	518	319
CA5	515	1097	887	2630	1167	561	284

1) Lire les données en `matlab`.
2) Calculer la matrice de covariance.
3) Calculer la moyenne et l'écart-type des données.