

# TABLE DES MATIERES

<b>PARTIE 1 PREALABLES A UN TRAITEMENT STATISTIQUE</b> .....	17
<b>CHAPITRE 1 UNE DEMARCHE SCIENTIFIQUE</b> .....	19
1. PREALABLE A UNE ANALYSE STATISTIQUE.....	19
2. SCHEMA D'UNE DEMARCHE SCIENTIFIQUE.....	21
3. D'UN BUT A UNE REALISATION.....	22
3.1 Triplet fondamental.....	22
3.2 D'une vision idéale à la réalité.....	24
3.3 Utilisation d'un modèle, sa confrontation à la réalité.....	26
3.4 Sélection des variables explicatives d'un modèle.....	27
3.5 Choix d'un modèle et validation de résultats.....	28
4. L'OBTENTION D'UN CORPUS DE DONNEES UTILISABLE.....	30
4.1 La croissance des corpus de données.....	30
4.2 Les systèmes de gestion de base de données.....	32
5. ORGANISATION DE L'OUVRAGE.....	33
<b>CHAPITRE 2 LES OUTILS DE REPRESENTATION D'UN ECHANTILLON</b> .....	37
1. STRUCTURE D'UN TABLEAU DE DONNEES.....	37
1.1 Questions préalables.....	37
1.2 La forme du tableau de données.....	38
1.3 Notion de type.....	38
1.4 Représentation d'un tableau de données.....	39
2. RESUMES UNIDIMENSIONNELS.....	40
2.1 Graphiques.....	41
2.2 Paramètres numériques classiques.....	44
2.3 De l'intérêt des transformations des variables.....	48
2.4 Estimation de la densité.....	49
2.5 Les variables qualitatives.....	52
3. RESUMES MULTIDIMENSIONNELS.....	54
3.1 Espace de représentation : point moyen $\bar{x}$ .....	54
3.2 Première transformation : matrice de dispersion.....	55
3.3 L'espace des observations $\mathbb{R}^p$ : regards sur les lignes, inertie.....	57
3.4 L'espace des variables $\mathbb{R}^n$ : regards sur les colonnes.....	59
3.5 Transformation linéaire d'un tableau de données.....	59
4. DECOMPOSITION D'UNE MATRICE DE DONNEES.....	60
4.1 Décomposition en Valeurs Singulières (DVS).....	60
4.2 Inverse généralisée d'une matrice.....	61
4.3 Exemples numériques.....	62
4.4 Décomposition en valeurs singulières du triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ .....	65
5. BILAN.....	66
PROBLEMES ET EXERCICES.....	67

<b>CHAPITRE 3 PRATIQUES UTILES AVANT TRAITEMENT</b> .....	69
1. TRANSFORMATIONS DES DONNEES .....	69
1.1 La famille de transformations de Box & Cox .....	69
1.2 Exemple des eaux minérales.....	70
2. RE-ECHANTILLONNAGE DES DONNEES .....	72
2.1 Autovalidation des résultats .....	72
2.2 Validation croisée.....	72
2.3 <i>Jackknife</i> .....	72
2.4 <i>Bootstrap</i> .....	75
2.5 Un exemple (Tabac) .....	75
2.6 Test de permutation .....	78
3. DONNEES SUSPECTES .....	80
3.1 Rappel du cas unidimensionnel .....	80
3.2 Cas multidimensionnel.....	83
4. LES DONNEES MANQUANTES SONT INEVITABLES.....	88
4.1 Que faire avec des données manquantes ? .....	88
4.2 Origine des données manquantes .....	89
5. QUELLE METHODE POUR QUEL MECANISME ? .....	90
5.1 Méthodes conventionnelles : méthodes d'imputation simples.....	90
5.2 Estimation par Maximum de vraisemblance.....	93
5.3 Un exemple (Eaux2010).....	94
5.4 L'imputation multiple.....	98
6. BILAN .....	102
PROBLEME ET EXERCICES .....	103
<b>PARTIE 2 ETUDE D'UN ECHANTILLON</b> .....	105
<b>CHAPITRE 4 REPRESENTATION D'UN ECHANTILLON PAR DES CARTES : ACP</b> .....	107
1. QUAND UTILISER UNE ACP ? .....	107
2. PRINCIPE DE L'ACP .....	108
2.1 Déformation d'un nuage de points par projection.....	108
2.2 Vocabulaire de l'ACP .....	109
2.3 Reconstitution de la matrice de données .....	110
3. INTERPRETATION ET QUALITE DES RESULTATS DE L'ACP .....	111
3.1 Quelques règles d'interprétation.....	111
3.2 Interprétation de la position des observations et des variables .....	114
4. EXEMPLE : CALCULS DE BASE.....	117
4.1 Nombre d'axes .....	117
4.2 Analyse des variables (Tab.5).....	118
4.3 Analyse des observations.....	119
4.4 Validation .....	122
4.5 Représentation simultanée : graphe <i>biplot</i> .....	125
5. ANALYSE D'UN TABLEAU DE DISTANCES.....	127
5.1 Distances euclidiennes .....	127
5.2 Distances non euclidiennes.....	128
6. EXEMPLES .....	129
6.1 Retour sur les eaux minérales : distances euclidiennes .....	129
6.2 Distances entre villes, distances non euclidiennes ( <i>capitales (ade4)</i> ).....	129
7. BILAN .....	131
PROBLEMES ET EXERCICES.....	132

**CHAPITRE 5 REPRESENTATION D'UN ECHANTILLON PAR DES CARTES :**

<b>AFC ET AFCM</b> .....	133
1. L'AFC .....	133
1.1 Etude globale d'un tableau de contingence .....	133
1.2 Eléments déduits d'un tableau de contingence .....	134
1.3 Profils et pondération .....	135
2. DISTANCES, METRIQUE ET ACP .....	137
2.1 Distance entre 2 profils ligne .....	137
2.2 Distance entre 2 profils colonne .....	137
2.3 ACP des deux nuages .....	137
3. INTERPRETATION ET QUALITE DES RESULTATS EN AFC .....	140
3.1 Choix des axes .....	140
3.2 Qualité de représentation $q/t$ et Contribution $ctr$ .....	140
3.3 Représentation graphique des profils .....	142
4. EXEMPLE : REPARTITION DES TACHES MENAGERES .....	143
4.1 Les différents résultats de l'AFC .....	143
4.2 Remarque : décomposition en valeurs singulières de $N$ .....	147
5. EXTENSION DE L'AFC : L'AFCM .....	148
5.1 Cas de deux classes .....	148
5.2 Extension à plus de deux classes .....	149
5.3 Exemple : préférences de consommateurs ( $PrefConsum$ ) .....	152
6. BILAN .....	156
6.1 Validation .....	156
6.2 Application de l'AFC à d'autres tableaux de données .....	156
PROBLEMES ET EXERCICES .....	157
<b>CHAPITRE 6 ANALYSE FACTORIELLE : LE MODELE FACTORIEL</b> .....	159
1. INTRODUCTION : MODELE FACTORIEL ORTHOGONAL .....	160
1.1 L'ACP peut-elle être considérée comme un modèle ? .....	160
1.2 Modèle factoriel .....	161
1.3 Non unicité des pondérations des facteurs .....	163
2. ESTIMATION DES PARAMETRES .....	163
2.1 Méthode des composantes principales .....	164
2.2 Méthode des facteurs principaux .....	165
2.3 Méthode du maximum de vraisemblance .....	166
2.4 Choix du nombre de facteurs communs .....	166
2.5 Estimation des scores des facteurs communs .....	167
3. NON UNICITE DE LA SOLUTION ET ROTATION DES FACTEURS .....	169
3.1 Rotation orthogonale .....	169
3.2 Rotation oblique .....	169
4. EXEMPLES .....	169
4.1 Exemple élémentaire .....	170
4.3 Consommation de drogue en milieu étudiant ( $usagedroque.cor$ ) .....	177
5. BILAN .....	182
5.1 Validité du modèle factoriel, comparaison ACP et AFCS .....	182
5.2 Logiciels .....	182
PROBLEMES ET EXERCICES .....	183

<b>CHAPITRE 7 REPRESENTATION D'UN ECHANTILLON PAR DES CLASSES</b> .....	185
1. QUAND UTILISER UNE METHODE DE CLASSIFICATION ? .....	185
2. IDEES GENERALES .....	186
2.1 Classes polythétiques et classes monothétiques .....	186
2.2 Mesures de ressemblance .....	187
2.3 Concepts courants en classification .....	189
2.4 Caractère combinatoire de la classification .....	189
3. CLASSIFICATION PAR PARTITION .....	189
3.1 Inertie inter-classe et inertie intra-classe.....	189
3.2 Regroupement d'observations autour de centres mobiles : méthode <i>k-means</i> .....	191
3.3 Exemple des eaux minérales (Eaux1).....	192
3.4 Comparaison de moyennes.....	200
4. CLASSIFICATION HIERARCHIQUE.....	201
4.1 Classification ascendante hiérarchique ou <i>CAH</i> .....	201
4.2 Classification descendante hiérarchique.....	209
4.3 Classification monothétique par division.....	211
5. MODELES DE MELANGE POUR LA CLASSIFICATION .....	215
5.1 Principes.....	215
5.2 Exemple (planete).....	217
6. CLASSIFICATION AVEC RECOUVREMENTS.....	218
7. CLASSIFICATION DE VARIABLES .....	219
8. BILAN .....	220
PROBLEMES ET EXERCICES.....	221
 <b>PARTIE 3 ETUDE DE DEUX GROUPES DE VARIABLES</b> .....	223
 <b>CHAPITRE 8 REGRESSION : LES BASES ET LES LIMITES</b> .....	225
1. QUESTIONS QUE PERMET D'ABORDER LA REGRESSION .....	225
2. MODELE LINEAIRE.....	227
2.1 La Régression linéaire multiple .....	227
2.2 Identification d'un « bon » modèle et validation .....	231
3. EXEMPLE : ETUDE DU RENDEMENT FROMAGER ( <i>RdtFromage</i> ) .....	238
4. REGRESSEURS QUALITATIFS : L'ANALYSE DE VARIANCE ( <i>ANOVA</i> ).....	246
4.1 Organiser les observations à traiter : la planification expérimentale .....	246
4.2 Quelques définitions .....	247
4.3 De l'analyse de variance à un facteur contrôlé à la régression .....	248
4.4 Analyse de variance à deux facteurs croisés .....	253
4.5 Analyse de variance et régression.....	257
4.6 Analyse de covariance ( <i>ANCOVA</i> ) .....	258
4.7 Données manquantes, déséquilibre que faire ?.....	261
5. UNE STRATEGIE POSSIBLE POUR ESTIMER LES MODELES DE REGRESSION .....	262
5.1 Premières approches, principes essentiels .....	262
5.2 Comment régler les problèmes avec les suppositions ? .....	263
6. BILAN .....	264
PROBLEMES ET EXERCICES.....	265

<b>CHAPITRE 9 LA COLINEARITE : DU DIAGNOSTIC AUX REMEDES .....</b>	<b>267</b>
1. EFFETS DE LA COLINEARITE ET DETECTION.....	267
1.1 Introduction .....	267
1.2 Détection et diagnostic .....	271
2. LA REGRESSION SUR COMPOSANTES PRINCIPALES (PCR) .....	276
2.1 La méthode .....	276
2.2 Application à la processionnaire du pin .....	278
3. REGRESSION <i>PLS</i> (Partial Least Squares).....	284
3.1 La méthode <i>PLS1</i> .....	285
3.2 Application de <i>PLS1</i> à la processionnaire du pin.....	288
4. REGRESSION BIAISEE OU PENALISEE.....	292
4.1 Régression <i>Ridge</i> ou pseudo-orthogonalisée .....	293
4.2 Régression <i>Lasso</i> .....	296
4.3 Régression pénalisée et sélection de variables .....	298
5. BILAN .....	299
PROBLEMES ET EXERCICES.....	301
<b>CHAPITRE 10 RELATIONS ENTRE DEUX GROUPES DE VARIABLES .....</b>	<b>303</b>
1. L'ANALYSE DES CORRELATIONS CANONIQUES (ACC).....	303
1.1 Principe et origine .....	303
1.2 Formulation classique .....	304
1.3 Autres présentations.....	306
2. PREMIERS ELEMENTS D'INTERPRETATION .....	308
2.1 Dimension de l'espace de représentation canonique .....	308
2.2 Outils de représentation et d'interprétation .....	309
3. EXEMPLE HISTORIQUE : MENSURATIONS SUR DES JUMEAUX.....	314
3.1 Description et ACC .....	314
3.2 Dimension de la représentation.....	316
3.3 Vision interne de l'espace canonique de chaque groupe.....	318
3.4 Relation entre les deux espaces canoniques.....	319
4. COMPLEMENTS ET EXTENSIONS .....	320
4.1 Ré-échantillonnage .....	320
4.2 Vérifications, prédiction .....	320
4.3 Extensions .....	320
5. AUTRES METHODES .....	321
5.1 Analyse procustéenne (AP) .....	321
5.2 Méthodes factorielles .....	325
6. ELEMENTS POUR L'ANALYSE DU CORPUS « SOL/BLÉ » .....	331
6.1 Le corpus de données Sol/Blé (1nESB162).....	331
6.2 Difficultés a priori.....	332
6.3 Quelques sorties utiles.....	332
7. BILAN .....	337
PROBLEMES ET EXERCICES.....	338

**PARTIE 4 ETUDE DE PLUSIEURS ECHANTILLONS .....339****Chapitre 11 DISCRIMINATION ET CLASSEMENT :**

<b>I - COMMENT DECRIRE LA SEPARATION DE CLASSES .....</b>	<b>341</b>
1. OBJECTIFS ET PARTICULARITES D'UNE ANALYSE DISCRIMINANTE .....	342
1.1 Quelques champs d'application de l'analyse discriminante .....	342
1.2 Les deux facettes de l'analyse discriminante .....	342
1.3 Deux fois rien, ça peut être beaucoup .....	342
2. APPROCHE GEOMETRIQUE : LES ASPECTS ESSENTIELS.....	344
2.1 Décomposition de la variabilité.....	344
2.2 Recherche d'un nouveau repère, sens du critère à maximiser .....	346
2.4 Les étapes du calcul, amphores crétoises (Amphore-a).....	349
2.5 Règles géométriques d'affectation ou règles de Fisher .....	354
3. EXEMPLE DE DEUX POPULATIONS ET VARIATIONS.....	358
3.1 Exemple avec deux populations, « Charolais*Zébus » (Chazeb-a).....	358
3.2 Une pratique critiquable mais utile, la sélection de variables .....	362
4. ASPECTS COMPLEMENTAIRES.....	363
4.1 Homoscédasticité.....	364
4.2 Dimension de la représentation.....	364
PROBLEMES ET EXERCICES.....	367

**Chapitre 12 DISCRIMINATION ET CLASSEMENT :**

<b>II - COMMENT AFFECTER DES OBSERVATIONS A DES CLASSES .....</b>	<b>369</b>
1. AFD AVEC REGLES D'AFFECTATION PROBABILISTES .....	369
1.1 Recherche d'une probabilité a posteriori.....	369
1.2 Loi Normale multidimensionnelle .....	370
1.3 Cas de deux populations .....	372
2. LA REGRESSION LOGISTIQUE BINAIRE .....	373
2.1 Principes de la régression logistique binaire .....	374
2.2 Estimation des paramètres : cas de la fonction de lien <i>logit</i> .....	374
2.3 Interprétation des coefficients $\beta_j$ dans le cas du lien <i>logit</i> .....	377
2.4 Evaluation : les pseudo $R^2$ .....	382
2.5 Sélection ou choix de modèles.....	383
2.6 Validation du modèle .....	384
3. OUTILS DE COMPARAISON ET DE QUALITE DES RESULTATS.....	386
3.1 Matrice de confusion .....	386
3.2 Sensibilité et spécificité.....	387
3.3 La qualité des résultats jugée par la courbe <i>ROC</i> .....	389
4. QUE CHOISIR : ANALYSE CLASSIQUE OU REGRESSION LOGISTIQUE ? .....	393
4.1. Les avantages comparés .....	394
4.2. Inconvénients et limites comparés .....	394
PROBLEMES ET EXERCICES.....	395

<b>PARTIE 5 AUTRES METHODES</b> .....	397
<b>CHAPITRE 13 ARBRES BINAIRES</b> .....	399
1. OBJECTIFS ET PARTICULARITES .....	399
1.1 L'apparence d'un arbre de décision .....	399
1.2 Comment se présente un arbre de décision .....	401
1.3 Un exemple simple d'arbre de décision .....	402
2. CRITERES DE CONSTRUCTION D'UN ARBRE DE DECISION BINAIRE .....	403
2.1 Questions préalables à la construction .....	403
2.2 Critère de division .....	404
2.3 Arbres de classement .....	405
2.4 Arbres de régression .....	407
3. COMMENT CHOISIR UN ARBRE ? .....	408
3.1 Définir la bonne taille de l'arbre .....	408
3.2 Traitement des données manquantes .....	408
3.3 Suppression des feuilles superflues .....	408
4. CONSTRUCTION D'ARBRES SOUS R .....	410
4.1 Arbre de régression .....	410
4.2 Arbre de classement .....	416
5. BILAN .....	422
PROBLEMES ET EXERCICES .....	423
<b>CHAPITRE 14 CONCLUSIONS ET PERSPECTIVES</b> .....	425
1. OUBLIS ET CONVICTIONS .....	425
1.1 Les oublis volontaires .....	425
1.2 Réfléchir, douter mais décider .....	426
1.3 Comprendre et expliquer .....	427
2. CHAMPS D'APPLICATION NON ABORDES .....	427
2.1 Fouille du web .....	427
2.2 Fouille de textes .....	428
2.3 Puces à ADN .....	429
2.4 Toxicologie et écotoxicologie .....	431
2.5 Mesures subjectives .....	431
3. ENVIRONNEMENTS NOUVEAUX, METHODES NOUVELLES .....	432
3.1 Grille informatique .....	432
3.2 Informatique en nuage .....	433
3.3 Machines à vecteur de support .....	433
3.4 Algorithmes génétiques .....	434
<b>ANNEXE LOGICIEL R ET DONNEES</b> .....	437
<b>BIBLIOGRAPHIE</b> .....	463
<b>INDEX</b> .....	475