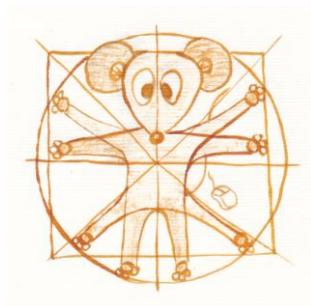


SSV- SSS ADD1

TD 5 : LDA

Robert Sabatier
Christelle Reynès



On se propose, d'analyser un jeu de données composé de 20 œuvres de Klimt, de Caravage, de Van Gogh et de Rembrandt (ordonnées ainsi dans le tableau). Afin d'effectuer une analyse statistique, on représente chaque image par son histogramme de couleurs. Cela consiste à partitionner l'espace des couleurs en k parties égales et, pour chaque image, calculer la proportion des pixels se trouvant dans la partie j , pour $j = \{1, \dots, k\}$. Ensuite, on associe à chaque image le vecteur numérique de taille k contenant les proportions des couleurs. Pour les 80 images, les vecteurs des histogrammes des couleurs avec $k=8$ et $k=64$ sont dans les fichiers painting8.txt et painting64.txt. Le fichier code_couleurs.xlsx donne la correspondance des couleurs pour chaque tableau.

Utilisation de la méthode LDA

1. Importer les données (format *matrix*) et créer le vecteur des groupes (format *factor*).
2. Explorer les données.
3. Appliquer une classification hiérarchique pour étudier les ressemblances entre échantillons. En déduire la nécessité ou non d'une transformation des données.
4. Pourquoi faut-il supprimer une variable avant de lancer la méthode LDA ?
5. Pour le jeu à 64 couleurs, éliminer les variables dont la variance est inférieure à 10^{-4} .
6. Appliquer la méthode LDA sur les deux jeux de données. Combien d'axes obtient-on ?
7. Quel pourcentage de bien classés obtient-on sur les données d'apprentissage (utilisation de la fonction *predict*) ?
8. Représenter le plan discriminant 1-2 pour les deux jeux de données. Repérer les erreurs. En déduire, le meilleur jeu de données pour la discrimination. On ne travaillera que sur ce jeu pour la suite des questions.
9. Représenter dans l'espace les trois axes discriminants (utilisation de la fonction *plot3d* de la library(*rgl*)).
10. Calculer les corrélations entre les valeurs des couleurs et les deux premiers axes discriminants et les représenter pour le premier plan. En déduire quelques couleurs candidates discriminantes et étudier leur distribution.
11. Utiliser la fonction *lda_CV* (qui nécessite d'avoir sourcé la fonction *constCV*) pour calculer le taux de bien classés de validation.