**TER : Scoring**

**II- Analyse discriminante**

1. Théorie de la méthode probabiliste

Le but de cette méthode est de prédire au mieux les valeurs d’une variable Y qualitative à K modalités, à partir de p variables explicatives X = (X1 , . . . , Xp ) quantitatives.

Dans cette section, nous allons classer un individu dans le groupe le plus "probable" et non dans le groupe le "plus proche" comme c'est le cas pour l'analyse discriminante géométrique. Pour ce faire, nous allons devoir faire des hypothèses probabilistes sur les données.

On suppose maintenant que les données sont issues d’une population regroupant des individus de K groupes prédéfinis différents G1; \_ \_ \_ ;Gk et que :

— Y est une variable aléatoire qui prend ses valeurs dans {1; \_ \_ \_ ;K}

— X = (X1; \_ \_ \_ ;Xp) est un vecteur de variables aléatoires réelles

On notera :

* pk = P(Y = k) la probabilité à priori d'appartenir au groupe Gk.
* fk = Rp 􀀀! [0; 1] la densité de X dans le groupe k
* wi le poids d’un individu xi
* wk =nk/n le poids de du groupe où nk représente le nombre d’individus dans le groupe
* gk= le centre de gravité du sous nuage formé par es individus du groupe
* Wk = (formule du cours RD4 de Labatte) : matrice de variance covariance du nuage Nk représentant le groupe
* W = (formule du cours RD4) : matrice de var\_cov intragroupe.

On supposera que l’on dispose d’un échantillon i.i.d. (X1; Y1); \_ \_ \_ ; (Xn; Yn) de même loi que (X; Y ).

* 1. Approche paramétrique : modèle Bayésien

Avec les notations précédentes, on obtient d’après la formule de Bayes que la probabilité qu’un individu x appartienne au groupe s’écrit :

On affecte alors l’individu x au groupe pour lequel la probabilité est la plus forte. (Règle du maximum a posteriori).

Or comme le dénominateur est constant pour un individu x donné, il suffit de déterminer le groupe pour lequel est le plus grand.

La fonction de densité introduite ici peut être déterminée soit :

1. Par des approches non paramétriques : on cherche à estimer directement à partir des données les densités (méthode des noyaux, des plus proches voisins)
2. Par des méthodes paramétriques : on suppose d’une forme paramétrique particulière et on estime les paramètres grâce à l’échantillon d’apprentissage.
   1. Modèle Bayésien avec méthode paramétrique

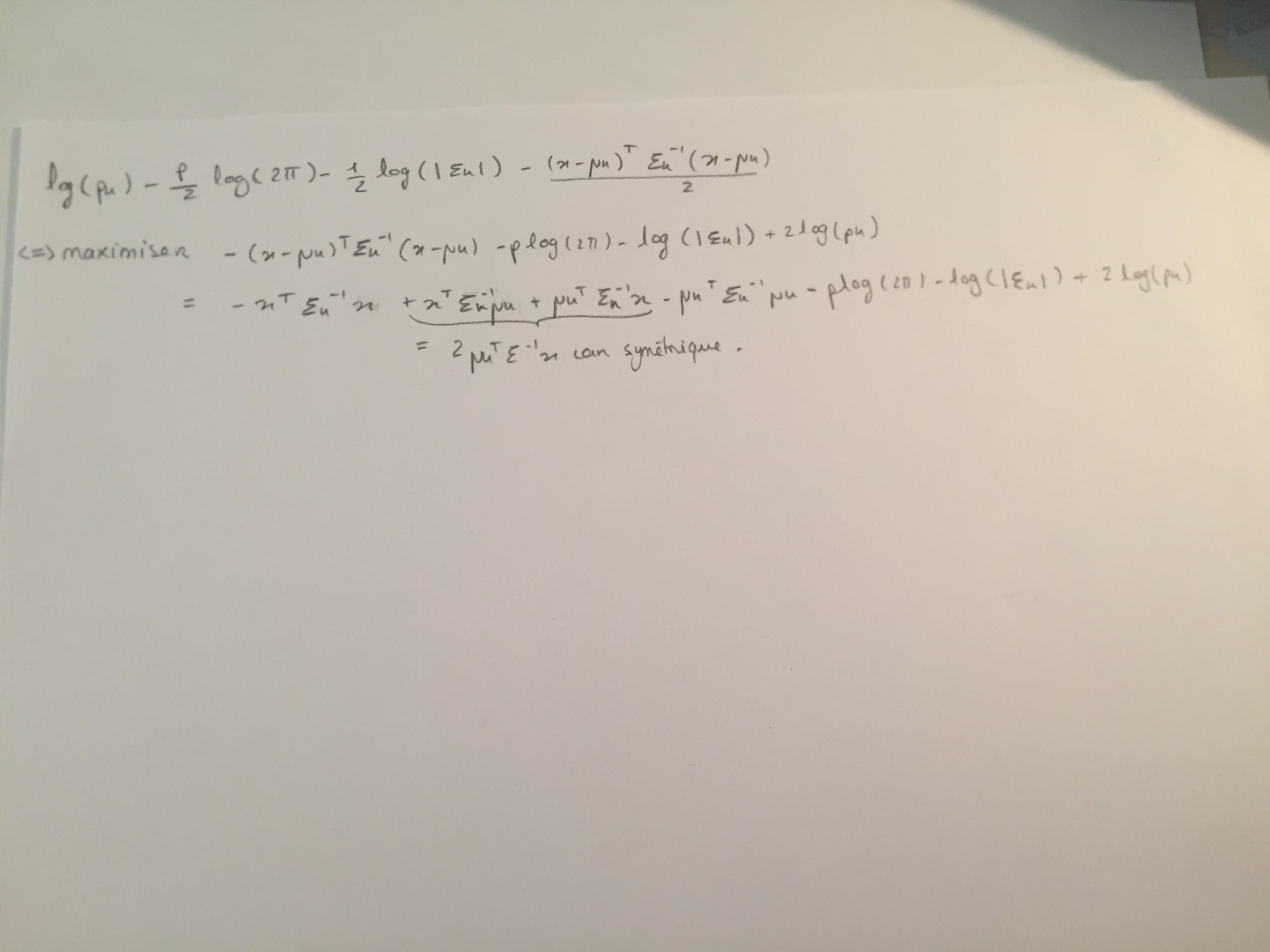
On considère dans cette section que X suit une loi normale dans chaque groupe. On se place donc dans le cas paramétrique Gaussien.

On a donc = formule cours RD4 page 3,

où appartient à Rp est le vecteur des moyennes théoriques et matrice des variances-covariances théoriques

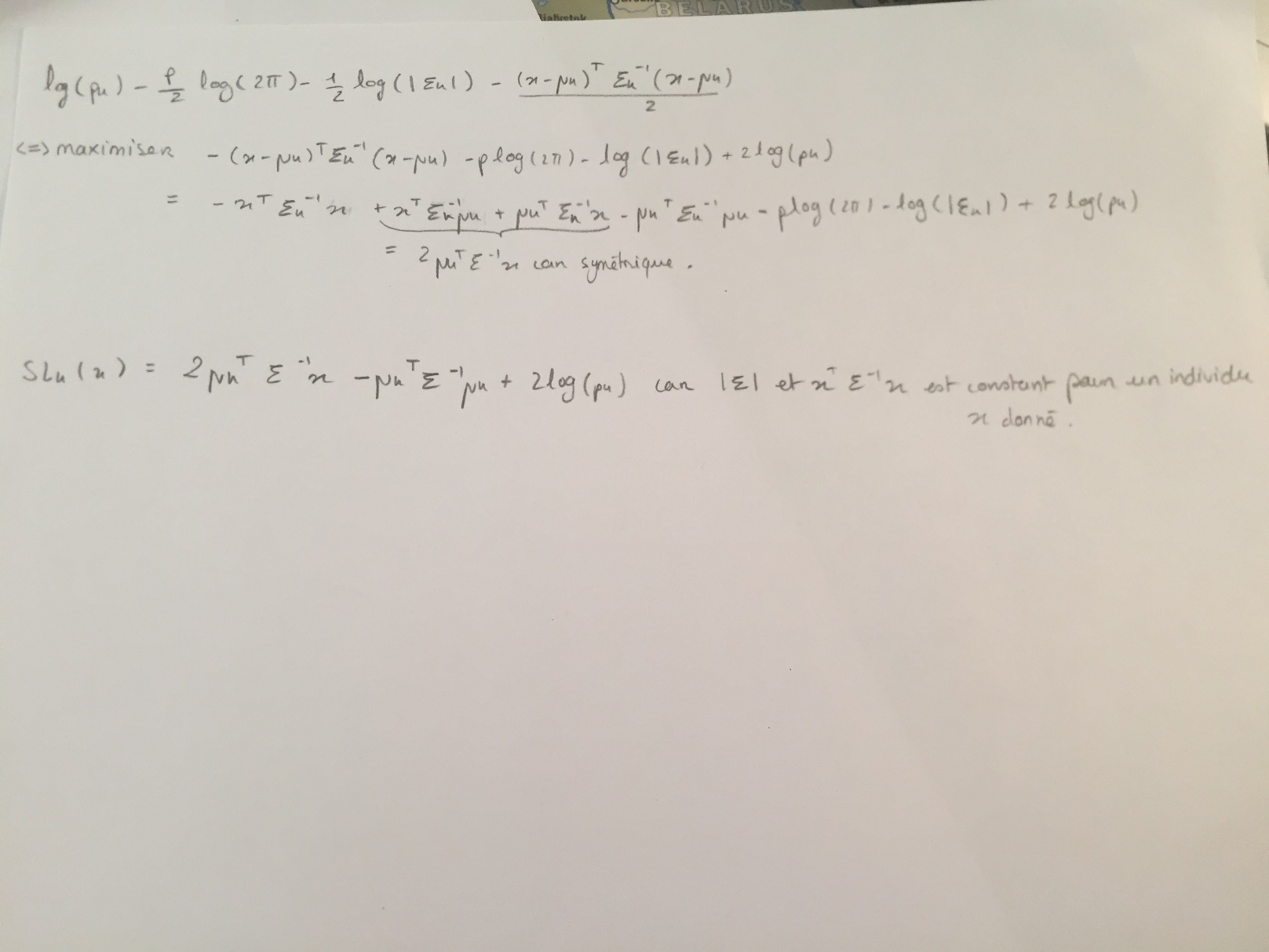
La règle de classement énoncée précédemment (aussi appelée règle de Bayes) revient donc à maximiser.

Par soucis pratique, on préfère maximiser le log de cette expression, c’est-à-dire maximiser :



* + 1. Analyse discriminante linéaire

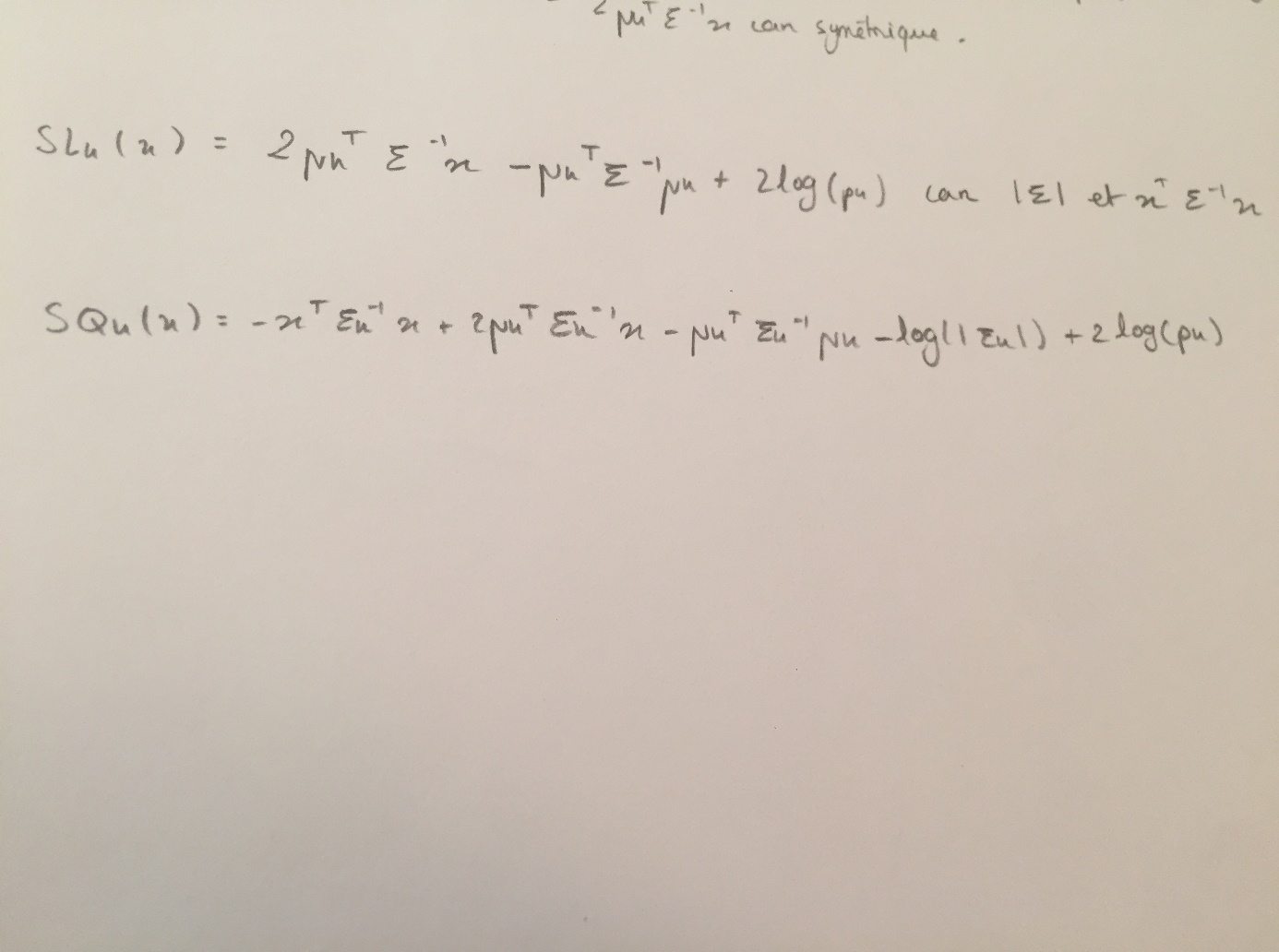
Dans le cas où les matrices de variance-covariance peuvent être supposées égales, on obtient le critère linéaire en x (d’affectation de l’individu x au groupe Gk) :



On affecte donc x au groupe Gk qui donne une valeur de SLk(x) maximale.

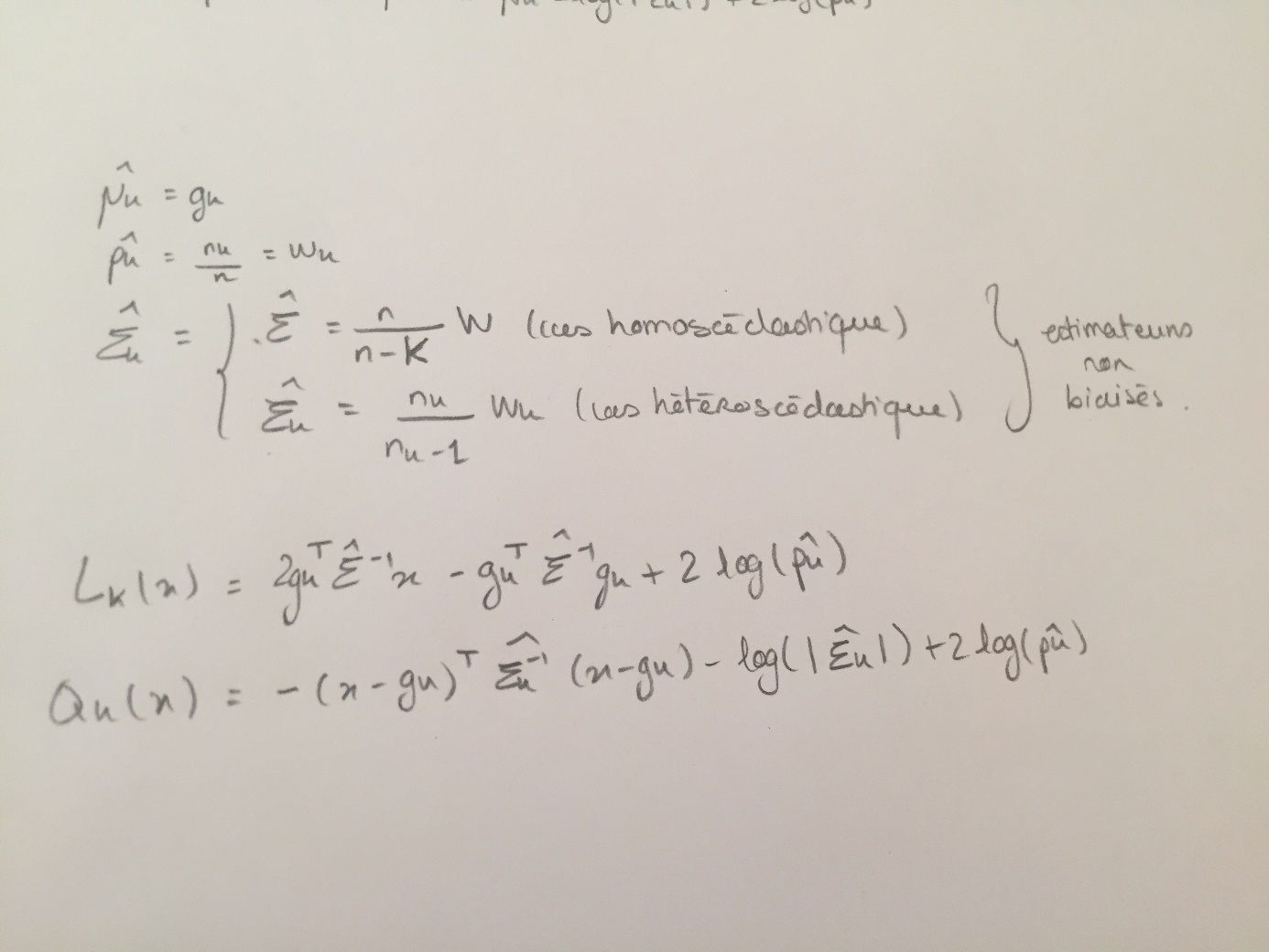
* + 1. Analyse discriminante quadratique

Dans le cas où les matrices de variance-covariance ne peuvent pas être supposées égales, on obtient le critère quadratique en x :



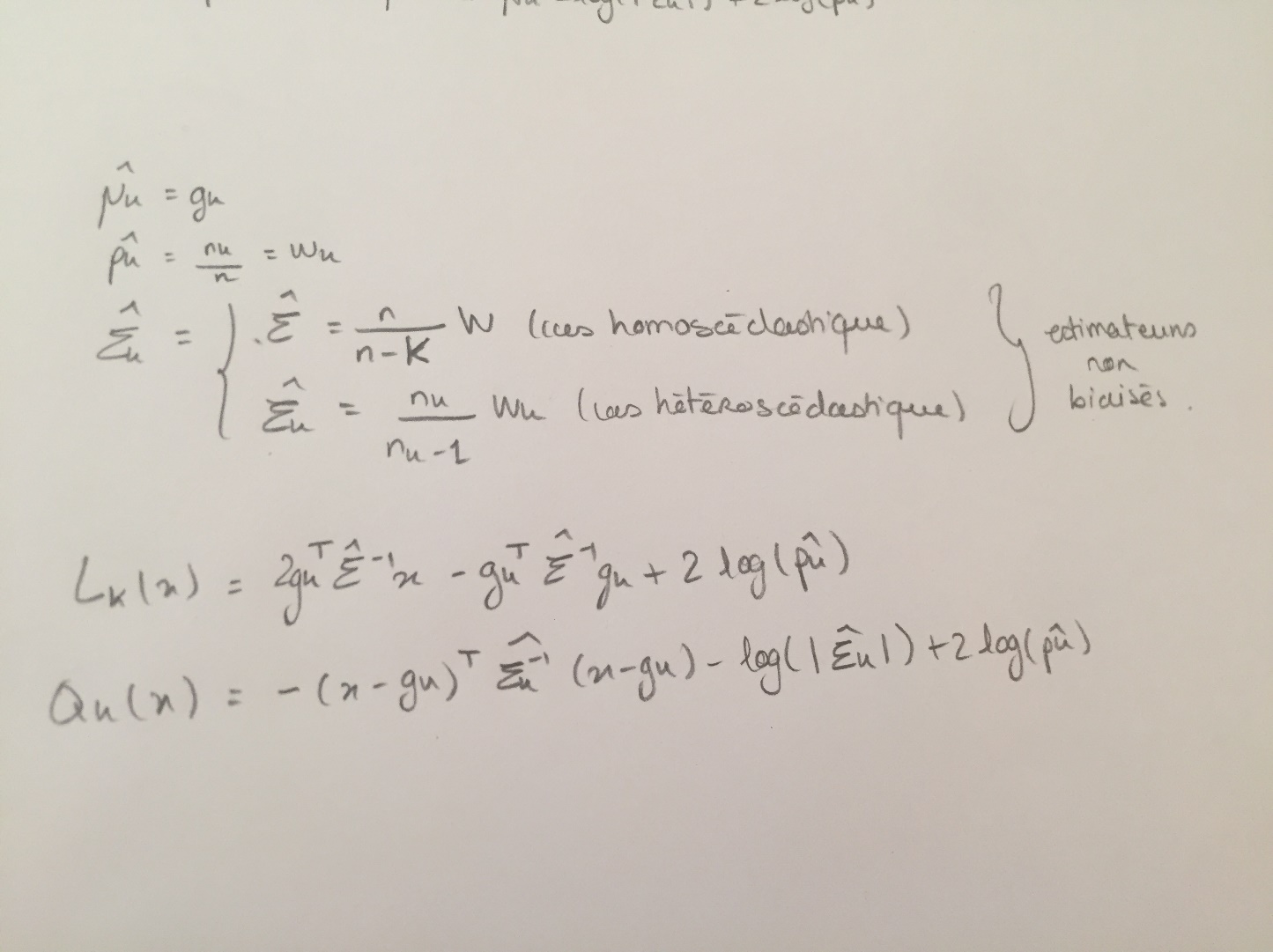
* 1. Estimation des paramètres

Le calcul de SLk(x) et SQk(x) n’est possible qu’en estimant les paramètres non connus en pratique. Ces paramètres peuvent être estimés par maximum de vraisemblance auquel cas on obtient :



* + 1. LDA

Dans le cas homoscédastique, on a alors :



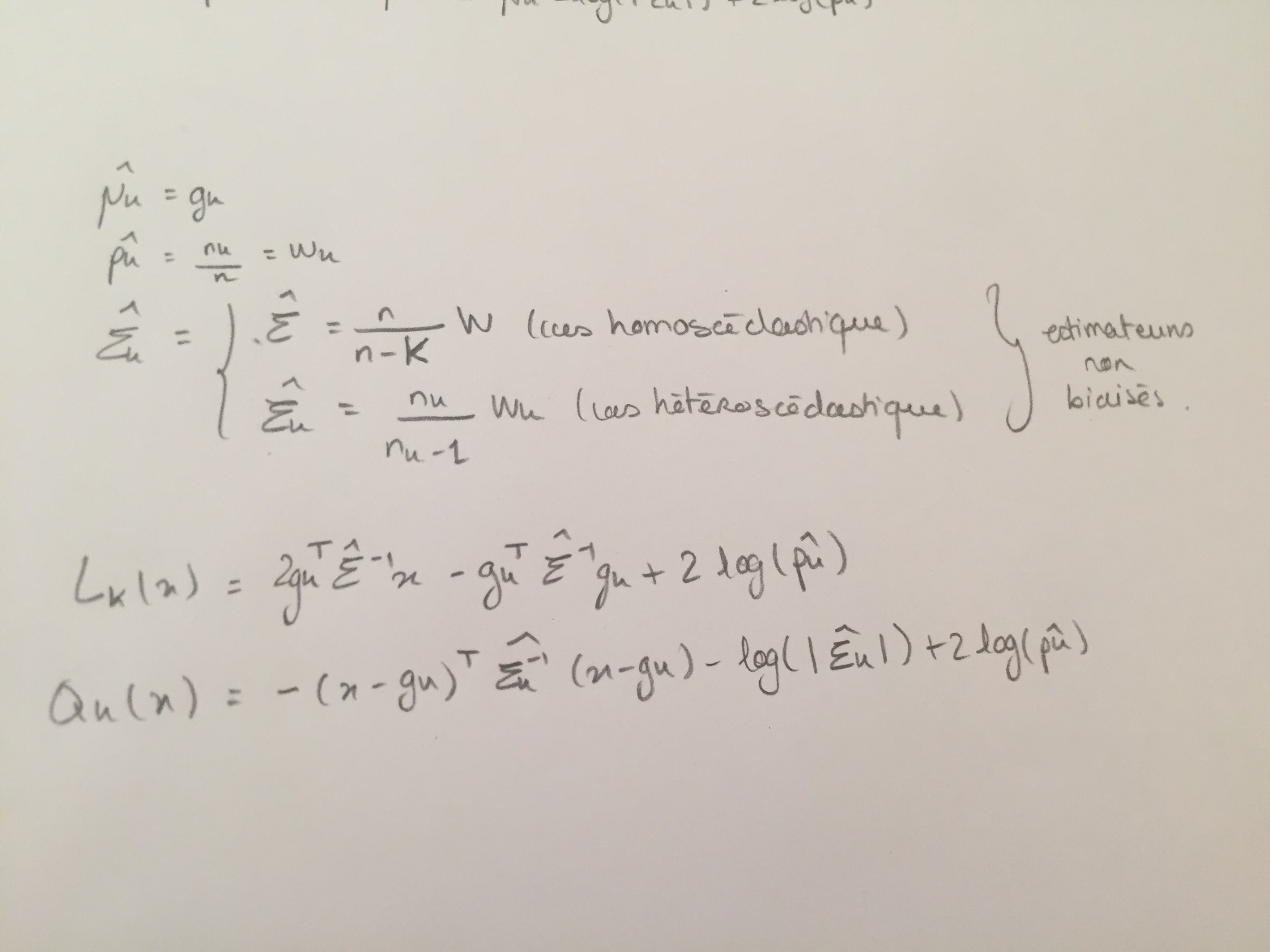
Cette fonction est appelée fonction linéaire discriminante du groupe Gk. Chaque fonction linéaire discriminante définie une fonction score et un nouvel individu sera donc affecté au groupe Gk pour lequel le score sera le plus élevé.

**Remarque** : On retrouve la fonction linéaire discriminante Lk(x) = (formule écrit2.pdf)

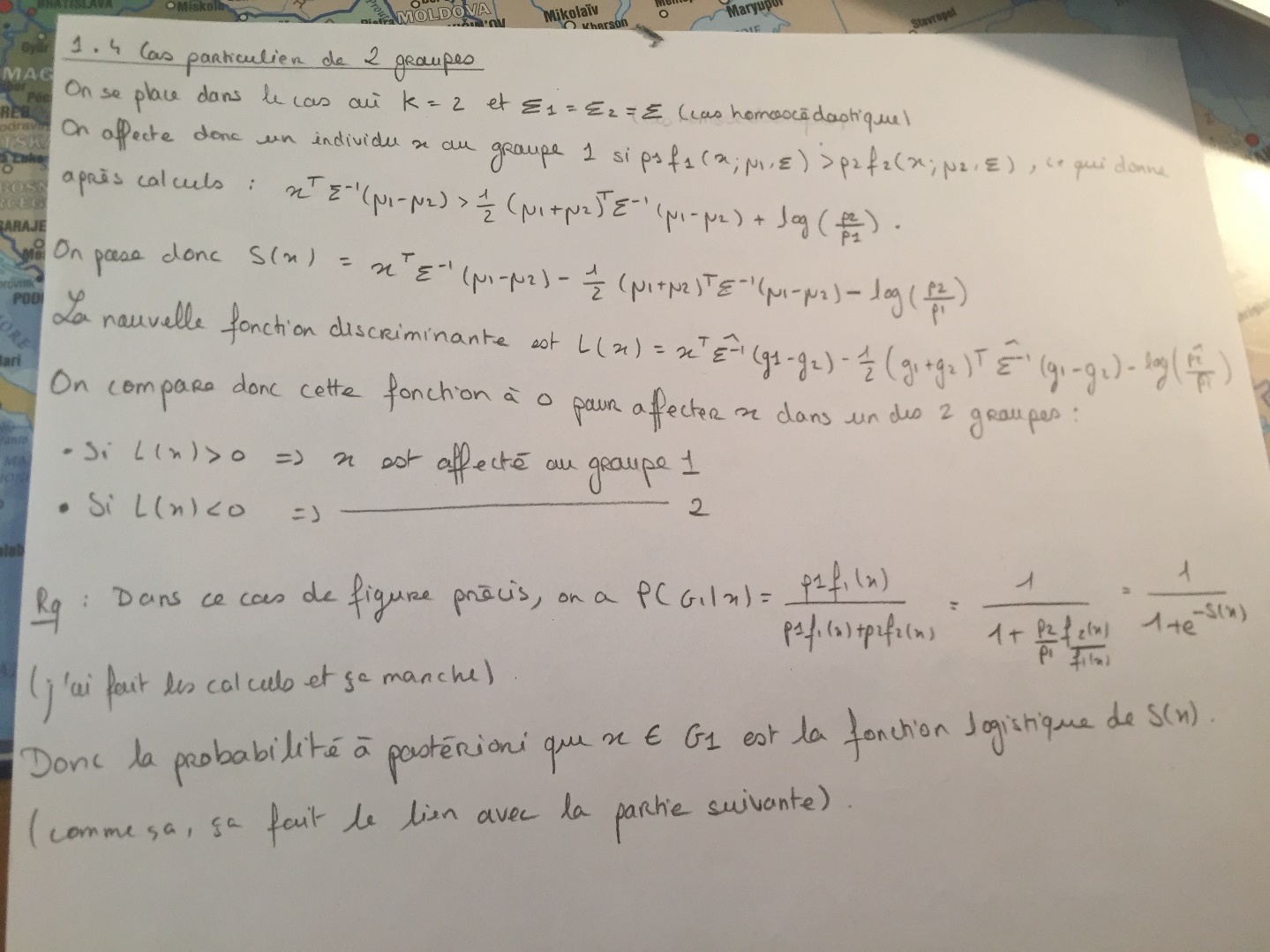
de l’analyse discriminante géométrique avec le terme ln(\_k) en plus. Dans le cas où l’on fait l’hypothèse d’égalité des probabilités à priori (p1 = \_ \_ \_ = pK), la règle de l’analyse discriminante linéaire (LDA) est équivalente à la méthode de l’analyse discriminante géométrique.

* + 1. QDA

Dans le cas hétéroscédastique, on a alors :



Cette fonction est appelée fonction quadratique discriminante du groupe Gk. Chaque fonction quadratique discriminante définie une fonction score et un nouvel individu sera donc affecté au groupe Gk pour lequel le score sera le plus élevé.



**Partie Protocole**

Sélection des variables

Le Lambda de Wilks est souvent utilisé dans les logiciels comme critère pour ne garder que les variables apportant de l’information sur l’appartenance ou non d’un individu à un groupe.

Le Lambda de Wilks est une approche paramétrique permettant de tester si plusieurs variables continues distinctes  sont liées à une variable qualitative Y à K ≥ 2 groupes, lorsqu’elles sont considérées avec leurs différentes interactions multivariées.

Les hypothèses d’utilisation de ce test sont:  suivent une loi normale et leur matrice de covariance respective sont égales (homoscédasticité).

La statistique du test du Lambda de Wilks se définie de la manière suivante :

Où W est la matrice de variance-covariance intragroupe et B la matrice de variance-covariance intergroupe.

Cette statistique de test suit une loi de Wilks à  degrés de liberté et l’hypothèse  est : « Indépendance entre  et  ».

Une variable a un bon pouvoir discriminant si la dispersion intra-groupe est faible et si la dispersion intergroupe est forte. Donc plus le Lambda de Wilks sera faible, plus la variable considérée est discriminante.

C’est ce critère qu’utilise la commande « greedy.wilks » de Rstudio que nous avons utilisée pour trouver les variables les plus discriminantes dans notre jeu de données et ainsi se focaliser sur un nombre de modèles plus réduit.