TP de Data Mining
Petit problème purement théorique

Maxime CHAMBREUIL
maxime.chambreuil@insa-rouen.fr

9 mars 2003

1 Explication

Contrairement à la semaine dernière, nous n'avons pas trouvé de fichier de données pour faire nos calculs. Nous avons donc généré 2 variables continues pour 50 individus virtuels à partir de paramètres connus :

$$\mu = \begin{pmatrix} 2 & 1 \end{pmatrix}$$

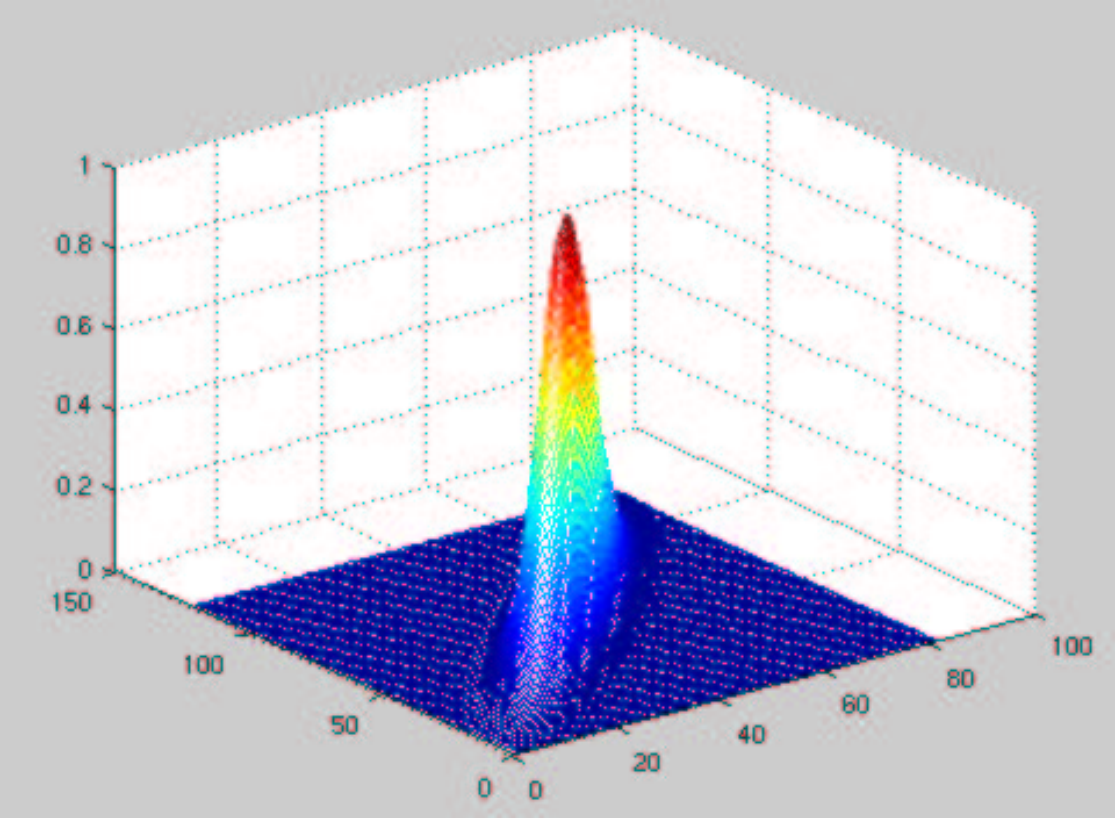
$$\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$$

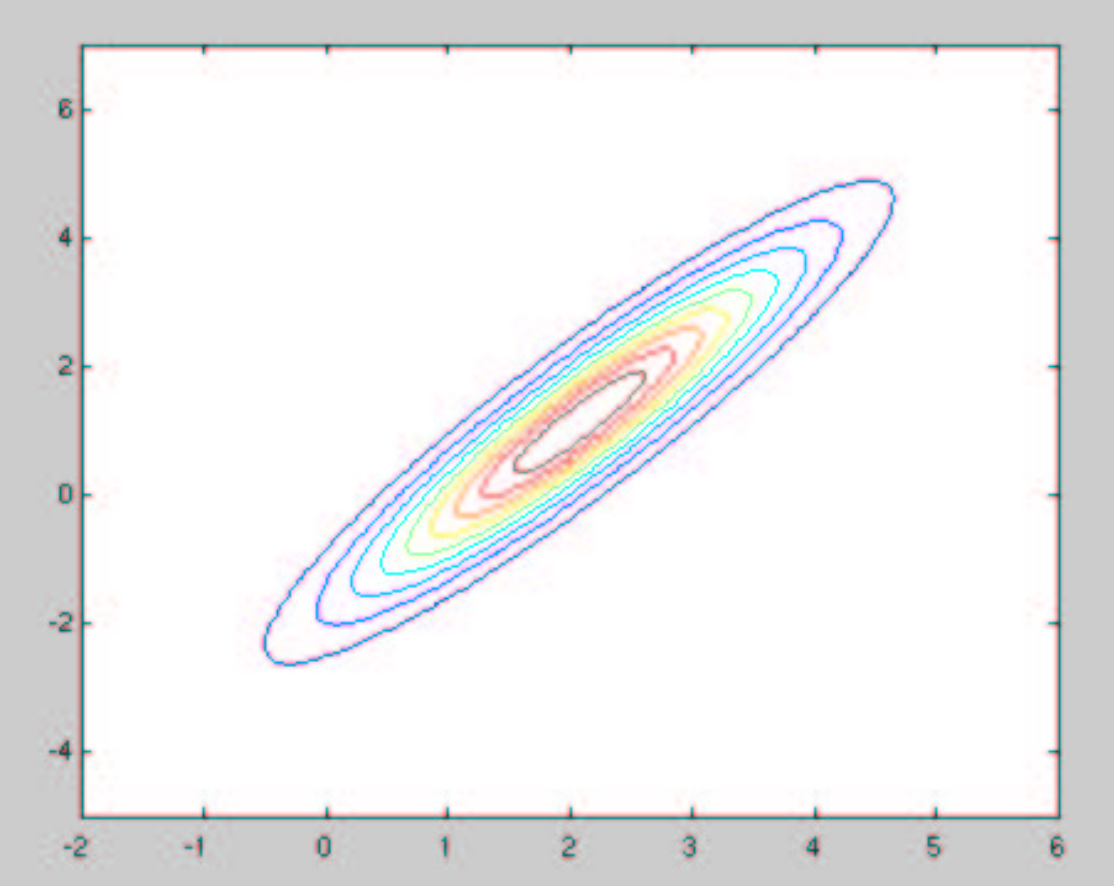
A partir de ses 50 réalisations, on a calculé la moyenne et la variance empirique pour estimer les paramètres :

$$m = \begin{pmatrix} 2.235568 & 1.330064 \end{pmatrix}$$

$$S = \begin{bmatrix} 1.295839 & 1.713674 \\ 1.713674 & 2.743403 \end{bmatrix}$$

On calcule ensuite la loi jointe à l'aide des paramètres estimés, puis on trace sa fonction de probabilités, selon nos 2 variables :


1
0.8
0.6
0.4
0.2
0
150
100
50
0
0
20
40
60
80
100

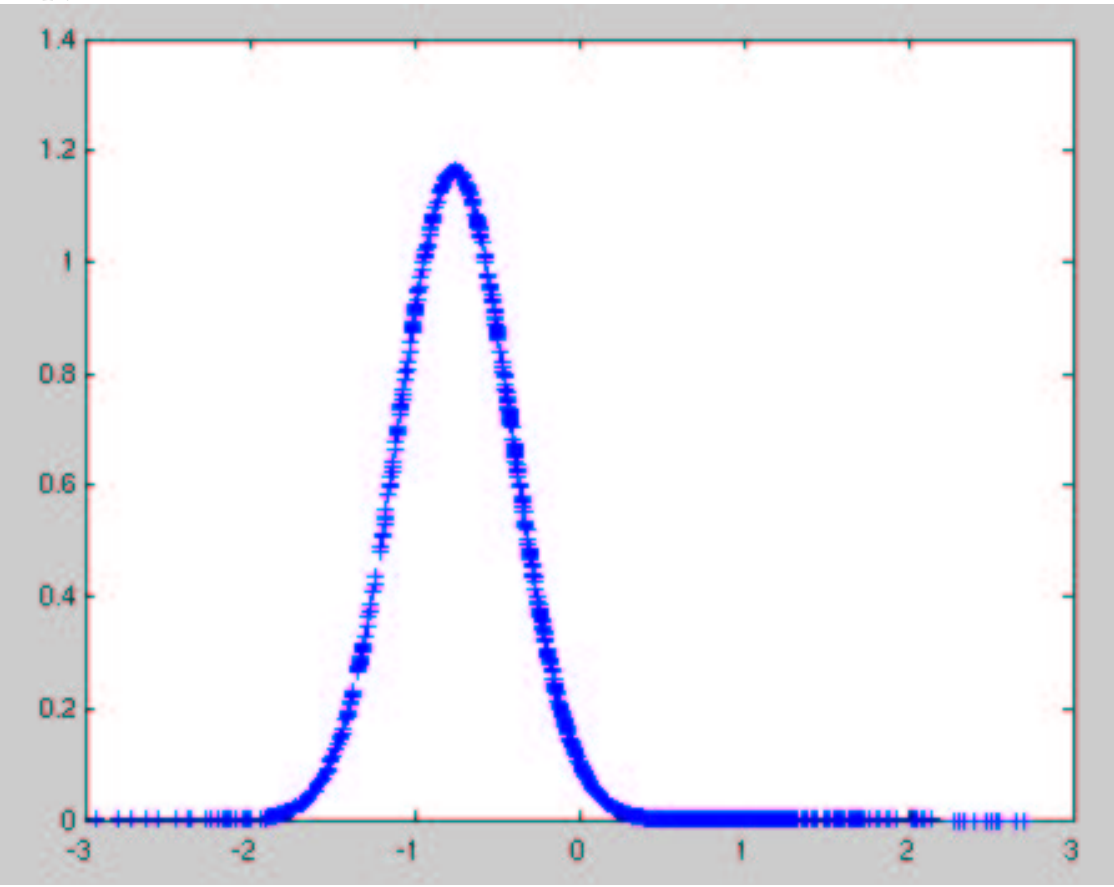

6
4
2
0
-2
-4
-2
-1
0
1
2
3
4
5
6

Grâce aux formules du cours, on a calculé les paramètres de la loi conditionnelle de x2 sachant x1 = -1 :

$$moyenneCondi = -0.8296$$

$$varianceCondi = 0.2870$$

Ainsi, nous avons pu tracer la loi de probabilité conditionnelle de x2 sachant x1 = - 1 :



2 Interprétation

On peut déjà dire qu'on a une bonne approximation des paramètres avec notre échantillon de 50 individus.

On remarque que le fait de traiter des données continues ne nous a pas poser de problème pour le calcul conditionnelle.

Comme dans le TP précédent, on va pouvoir générer des nouveaux individus à partir de nos calculs et présupposer la valeur d'une variable en fonction de l'autre, grâce aux calculs de probabilités conditionnelles.

Ainsi, on peut imaginer des comparaisons de populations : si on considère le poids et la taille, on va pouvoir en déduire que pour une taille donnée, un américain aurait plus d'embonpoint qu'un français, par exemple.

3 Conclusion

Ce TP m'a permis de découvrir et d'utiliser les fonctions d'affichage : meshgrid, mesh, contour, plot3. Ainsi j'ai pu me familiariser avec les dimensions de leurs paramètres. A l'avenir, je serais capable de confirmer les dimensions d'une donnée en la représentant graphiquement à l'aide de la fonction adéquate.

Je me suis rendu compte aussi que des estimations n'ont pas forcément besoin d'un trés grand nombre de valeur pour être proche de leur valeur asymptotique. Je ne pensais pas que 50 individus suffiraient mais plutôt 1 000 ou 10 000.