## Exercici 1: reducció de la dimensionalitat

### Apartat A

**Genereu un conjunt de dades sintètiques amb N=1000 observacions i 4 variables. La primera variable (x1) ha de seguir una distribució normal univariada amb mitjana 0.5 i desviació típica 0.2, és a dir x1~N(0.5, 0.1). La segona variables serà x2~N(4.5, 0.6). La tercera variable serà una combinació lineal de la primera x3 = 3\*x1 + 1. La quarta, una combinació no lineal de les dues primeres x4 = 2\*x12 + x2. Finalment, construïu una matriu de dades A de tamany Nx4 amb les variables x1, x2, x3, x4.**

### Apartat B

**Per cada combinació de variables, representeu les dades gràficament com a núvols de punts bidimensionals (x1 vs x2, etc.). Podeu fer servir la funció scatter de la llibreria matplotlib:**[**http://matplotlib.org/api/pyplot\_api.html?**](http://matplotlib.org/api/pyplot_api.html?)**highlight=scatter#matplotlib.pyplot.scatter. Comenteu breument les gràfiques obtingudes i la forma del núvol de punts tenint en compte les propietats de les dades.**

### Apartat C

**Apliqueu PCA a les dades anteriors i comenteu els resultats obtinguts tenint en compte les propietats de les dades. La funció PCA de scikit-learn està descrita a  http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html .**

**Quantes variables descorrelacionades hi han? Quants components principals són necessaris per explicar un 95% de la variabilitat de les dades?**

### Apartat D

**Compareu els resultats obtinguts als apartats anteriors amb els que s'obtindrien en cas que les 4 variables estiguessin distribuïdes amb distribucions independents x~N(1, 0.1). Justifiqueu raonadament els resultats obtinguts.**

## Exercici 2

### Apartat A

**Genereu un conjunt de dades bidimensional *A1* amb N=2000 observacions amb una distribució normal amb mitjana [5, -4] i matriu de covariança [2, -1; -1, 2], i un altre conjunt *A2* amb mitjana [1, -3] i matriu de covariança [1, 1.5; 1.5, 3]. Feu servir la següent funció de la llibreria *numpy*:**

**http://docs.scipy.org/doc/numpy/reference/generated/numpy.random.multivaria te\_normal.html**

### Apartat B

**Representeu les dades de tots dos conjunts en forma de núvol de punts en un únic gràfic. Feu servir simbols de color diferent per representar les dades de cada conjunt.**

### Apartat C

**Construïu un conjunt de dades d'entrenament (*training*) amb les 1000 primeres observacions de cada conjunt de dades i un conjunt de prova (*test*) amb les altres 1000. Feu servir les següents eines de classificació disponibles a les llibreries scikit-learn per dissenyar un sistema de classificació automàtica dels conjunts A1 i A2 a partir de les dades de test.**

**Naïve Bayes: Funció GaussianNB del paquet sklearn.naive\_bayes.**

**http://scikit- learn.org/stable/modules/generated/sklearn.naive\_bayes.GaussianNB.html**

**Anàlisi de Discriminants Lineal: Funció LDA del paquet sklearn.lda.**

**http://scikit-learn.org/stable/modules/generated/sklearn.lda.LDA.html**

**En cada cas, determineu el percentatge d'encerts i errors de classificació sobre el conjunt de validació.**

### Apartat D

**Repetiu l'anàlisi anterior quan la mitjana del primer conjunt de dades *A1* és [2, -4] en comptes de [5, -4]. Comenteu raonadament els resultats obtinguts i justifiqueu les diferències observades tenint en compte les característiques de les dades sintètiques utilitzades.**