Quelques éléments de statistique

B. Michel

Ecole Centrale de Nantes

Statistical Learning

Outline

Expérience statistique

- Une **expérience statistique** est la donnée d'un objet aléatoire X à valeurs dans un espace mesurable muni d'une famille de loi $(P_\theta)_{\theta\in\Theta}$, supposée contenir la loi de la variable aléatoire X.
- Par exemple, X peut être une v.a. continue ou discrète.
- L'ensemble des lois $(P_\theta)_{\theta\in\Theta}$ est connu, mais **le paramètre** θ de la "vraie" loi est inconnu. La démarche statistique consiste à donner de l'information sur le paramètre θ en s'appuyant sur des observations du phénomène X.
- Modèle paramétrique : $\Theta \subset \mathbb{R}^d$ et θ est le paramètre du modèle.
- On appelle n-**échantillon** la donnée d'un n-uplet $X = (X_1, \ldots, X_n)$ de v.a. indépendantes et de même loi p_θ.

Examples

- **Pile ou face.** Les résultats de n lancés consécutifs $X_1, \ldots, X_n$ d'une même pièce sont modélisés par n réalisations d'une même loi de Bernoulli de paramètre inconnu p.
  On considère ici les lois de Bernoulli $((\mu_p)^{\otimes n})_{p\in[0,1]}$ sur l'espace $\{0,1\}^n$.

- **Tailles.** On mesure la taille de n individus et on suppose que dans la population considérée la taille des individus suit une loi gaussienne de paramètres inconnus.
  L'expérience statistique est modélisée par un modèle gaussien unidimensionnel sur $\mathbb{R}^n$ muni des lois gaussiennes $\left((\nu_{\mu,\sigma^2})^{\otimes n}\right)_{\mu\in\mathbb{R},\sigma>0}$ de moyenne μ et de variance σ^2.

Outline

Estimateur

- On considère une expérience statistique X dans un modèle statistique paramétrique composé des lois $(P_\theta)_{\theta \in \Theta}$.
- Objectif : estimer une quantité $g(\theta)$ à partir de l'observation X, où g une application de Θ dans $\mathbb{R}^q$.
  Ex: $g(\mu, \sigma^2) = \mu$ dans le modèle gaussien précédent.
- Souvent on veut estimer θ lui-même et dans ce cas $g = Id$.

Definition

- *On appelle statistique toute variable (ou vecteur) aléatoire T à valeurs dans $\mathbb{R}^q$ qui est fonction de X :*

$$T = h(X)$$

  (où h est une application mesurable.)
- *Un estimateur $\hat{g}$ de $g(\theta)$ est une statistique à valeurs dans $g(\Theta)$, qui ne dépend pas de θ.*

Estimateurs : exemples

Dans le jeu du pile ou face :

- La quantité $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur de p qui semble pertinent (la loi des grands nombres s'applique...)
- La quantité $S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ est un estimateur de la variance $p(1-p)$.

Estimateurs : exemples

Dans le jeu du pile ou face :

- La quantité $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur de p qui semble pertinent (la loi des grands nombres s'applique...)
- La quantité $S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ est un estimateur de la variance $p(1-p)$.

mais :

- $\hat{p} = 0$ est aussi un estimateur (idiot) de p.

⚠ La définition d'un estimateur ne nous dit rien sur la qualité de l'estimation.

Fonction de vraisemblance (cas iid)

- On observe un n échantillon $X_1, \ldots, X_n$ sur $\mathbb{R}$ de loi commune p_θ de densité s_θ.
- La fonction de *vraisemblance* (*Likelihood* en anglais) pour le modèle paramétrique $(p_\theta)_{\theta \in \Theta}$ est la fonction définie sur Θ par

$$L_n(\theta) := \prod_{i=1}^{n} s_\theta(X_i)$$

- On considère aussi la fonction log-vraisemblance :

$$\ell_n : \theta \mapsto \ln(L_n(\theta)).$$

La méthode du maximum de vraisemblance (EMV)

- Méthode du maximum de vraisemblance : estimer le paramètre θ en choisissant le paramètre “le plus vraisemblable” pour l'échantillon disponible : i.e. trouver $\hat{\theta}_{\text{EMV}}$ pour lequel la vraisemblance $L_n(\theta)$ est maximale.
- On dit qu'un estimateur $\hat{\theta}$ est un estimateur du maximum de vraisemblance s'il vérifie

$$\hat{\theta} \in \text{argmax}_{\theta \in \Theta} \, L_n(\theta).$$

- Beaucoup de méthodes en Machine Learning reposent sur l'EMV.

EMV : exemples dans le cas continu

- Pour la famille des lois exponentielles, on vérifie que la fonction de vraisemblance vaut $L_n(\lambda) = \lambda^n \prod_{i=1,\dots,n} \exp(-\lambda X_i) \mathbb{1}_{\mathbb{R}^+}(X_i)$ et les X_i sont positifs p.s. On vérifie que $\ell_n(\lambda) = n \ln \lambda - \lambda \sum_{i=1,\dots,n} X_i$ et que $\hat{\lambda} = 1/\bar{X}_n$ est un EMV de λ.

- Pour un échantillon iid de variables aléatoires gaussiennes $\mathcal{N}(\mu, \sigma^2)$, on vérifie que $\bar{X}_n$ et $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ sont des EMV de μ et σ^2.

- Pour une loi Uniforme sur $[0, \theta]$, l'EMV vaut $\hat{\theta} = X_{(n)}$.

- Pour le modèle de loi Uniforme sur $[\theta - 1, \theta + 1]$, l'EMV n'est pas unique !

EMV : exemple dans le cas discret

- Dans le cas de variables aléatoires discrètes, la loi de référence à considérer est la mesure de comptage.

- Exemple : si $X_i \sim \mathcal{P}(\lambda)$, avec $\lambda > 0$, alors pour tout k on a $P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ et la densité de la loi de Poisson par rapport à la mesure de comptage sur $\mathbb{N}$ vaut donc

$$f_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

  pour tout $k \in \mathbb{N}$.

- La log-vraisemblance pour un échantillon $X_1 \ldots, X_n$ de variables aléatoires de lois de Poisson de paramètre λ vaut

$$\ell_n = -n(\lambda - \bar{X}_n \ln \lambda) - \sum_{i=1}^{n} \ln(X_i!),$$

- On vérifie alors que $\hat{\lambda} := \bar{X}_n$ dès que $\bar{X}_n > 0$.

Outline

Données Ozone

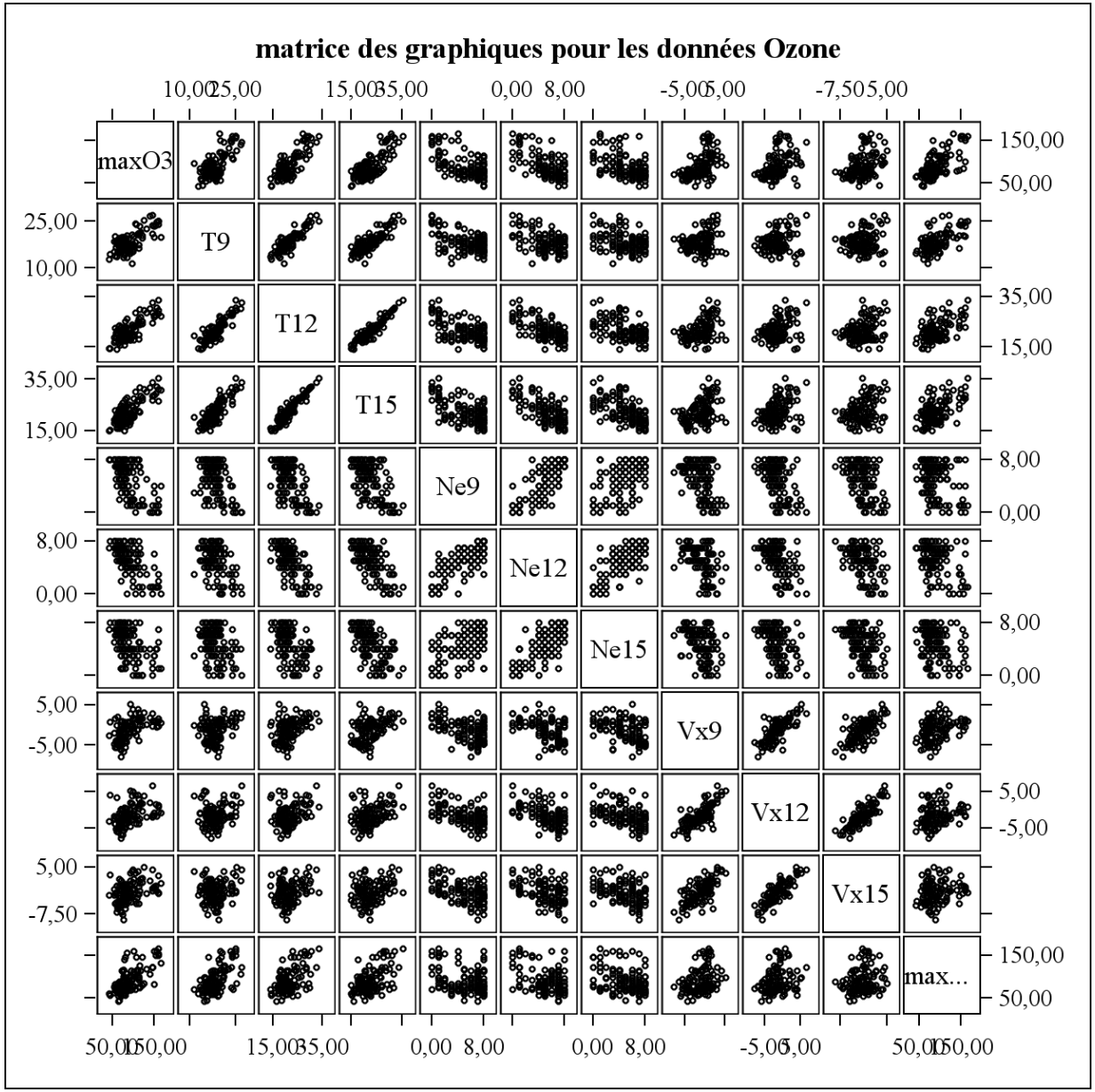
Jeu de données : niveaux de pollution enregistrés sur 112 journées de l'été 2001 à Rennes [1].

On souhaite expliquer la quantité d'ozone maxO3 par les mesures suivantes :

- T9, T12, T15 : température à 9h, 12h, 15h ;
- Ne9, N12, N15 : nébulosité à 9h, 12h, 15h ;
- Vx9, Vx12, VX15 : vitesse du vent sur un axe Est-Ouest à 9h, 12h, 15h ;
- MaxO3v : ozone mesurée la veille.

[1] exemple tiré de l'ouvrage *Statistiques avec R, P.A. Cornillon et al., PUR 2008*

Matrice du nuage de ces variables



Régression linéaire multiple

Dans le modèle de **régression multiple**, on suppose que l'espérance de Y est une fonction linéaire des variables explicatives x_j :

$$Y_i = \mathbb{E}Y_i + \varepsilon_i = a_0 + a_1 x_{i,1} + \cdots + a_{p-1} x_{i,p-1} + \varepsilon_i, \quad i = 1 \ldots n$$

où

- Y est la variable dite *dépendante* (à expliquer),
- $x_1, \ldots, x_{p-1}$ sont les variables explicatives ou regresseurs, supposées déterministes (design fixe)
  Remarque : la i-ème observation est le vecteur $(1, x_{i,1}, \ldots, x_{i,p-1})'$
- ε est le vecteur des erreurs, ces erreurs sont supposées indépendantes et de même loi centrée de variance σ^2.
- $\theta = (a_0, \ldots, a_{p-1})' \in \mathbb{R}^p$ et $\sigma > 0$ sont les paramètres du modèle. Lorsque $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ on parle de régression linéaire multiple gaussienne.

Écriture matricielle du modèle de régression linéaire multiple

Il est possible d'écrire ce modèle sous la forme matricielle suivante :

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p-1} \end{bmatrix}}_{\boldsymbol{X}=[\boldsymbol{e},x_1,\dots,x_{p-1}]} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{p-1} \end{bmatrix}}_{\boldsymbol{\theta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\boldsymbol{\varepsilon}}$$

$$Y = \boldsymbol{X}\boldsymbol{\theta} + \varepsilon$$

Modélisation de Y via une représentation "simple" de son espérance.

Problème des moindres carrés

Pour choisir θ on minimise la distance euclidienne entre Y et $\boldsymbol{X}\theta$:

Definition

- *On appelle estimateur des moindres carrés de $\boldsymbol{X}\theta$ un estimateur $\widehat{\boldsymbol{X}\theta}$ de $\boldsymbol{X}\theta$ qui minimise $\|Y - \widehat{\boldsymbol{X}\theta}\|^2$.*
- *On appelle estimateur des moindres carrés de θ un estimateur $\hat{\theta}$ qui minimise $\|Y - \boldsymbol{X}\hat{\theta}\|^2$.*

Remarque : un estimateur des moindres carrés de $\widehat{\boldsymbol{X}\theta}$ est aussi un minimiseur du risque empirique (ERM) pour la perte ℓ_2 dans l'espace des prédicteurs qui s'écrivent comme des combinaisons linéaires des variables explicatives.

Solution des moindres carrés

Proposition

- *L'estimateur des moindres carrés de $\boldsymbol{X}\theta$ est unique et vérifie*

$$\widehat{\boldsymbol{X}\theta}_{MC} = Argmin\{\|Y - u\|^2 \mid u \in Im(\boldsymbol{X})\} = \mathbf{P}_{Im(\boldsymbol{X})} Y.$$

- *θ minimise $\|Y - \boldsymbol{X}\theta\|^2$ si et seulement si θ est solution de*

$$\boldsymbol{X}'\boldsymbol{X}\,\theta = \boldsymbol{X}'Y. \tag{1}$$

Les équations (1) sont appelées **équations normales** du modèle linéaire.

Cas régulier

Lemma

Si $\boldsymbol{X}$ est injective (cas régulier) alors $\boldsymbol{X}'\boldsymbol{X}$ est inversible et il existe une unique solution aux équations normales :

$$\hat{\theta}_{MC} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Y.$$

On peut aussi vérifier que la matrice de projection sur l'image de $\boldsymbol{X}$ (aussi appelé *hat matrix* ou *hat operator*) vérifie :

$$\mathbf{P}_{\mathrm{Im}(\boldsymbol{X})} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

et on a bien que

$$\boldsymbol{X}\hat{\theta}_{\mathrm{MC}} = \widehat{\boldsymbol{X}\theta}_{\mathrm{MC}}.$$

Vecteur des valeurs prédites et MSE

- On définit le vecteur des valeurs prédites par (cas régulier)

$$\hat{Y} := \boldsymbol{X}\widehat{\theta}.$$

  Cette quantité $\hat{Y}$ peut être vue comme un estimateur de $\mathbb{E}Y$, ou comme un prédicteur du vecteur Y.

- Dans le cas régulier, au point de design x, le prédicteur vaut

$$\hat{y}(x) = x'\widehat{\theta}.$$

Vecteur des valeurs prédites et MSE

- On définit le vecteur des valeurs prédites par (cas régulier)

$$\hat{Y} := \boldsymbol{X}\widehat{\theta}.$$

  Cette quantité $\hat{Y}$ peut être vue comme un estimateur de $\mathbb{E}Y$, ou comme un prédicteur du vecteur Y.

- Dans le cas régulier, au point de design x, le prédicteur vaut

$$\hat{y}(x) = x'\widehat{\theta}.$$

- L'erreur moyenne quadratique (Mean Squared Error) vaut

$$MSE := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

  Cette quantité est égale au risque empirique pour la perte ℓ_2 du prédicteur linéaire $\hat{y}(x) = x'\hat{\theta}$.

Ce qu'il faut retenir de la régression linéaire

- Méthode standard et populaire pour la régression.
- Fournit un prédicteur de type linéaire : $\hat{y}(x) = x'\hat{\theta}$.
- Contrairement à de nombreuses méthodes plus complexes du ML, une description statistique fine de la régression linéaire est possible (non présentée dans cet exposé).
- Extension à la classification : régression logistique.
- Extension pour la grande dimension : le Lasso et ses variantes.

Outline

Espérance et variance de vecteurs aléatoires

- Pour un vecteur aléatoire X à valeur dans $\mathbb{R}^d$, l'espérance $\mathbb{E}X$ est le vecteur des espérances de ses composantes :

$$\mathbb{E}X = (\mathbb{E}X_1, \ldots, \mathbb{E}X_d)'.$$

- Pour deux vecteurs aléatoires X et Y à valeurs respectivement dans $\mathbb{R}^d$ et $\mathbb{R}^q$, la matrice de variance-covariance $\mathrm{Cov}(X, Y)$ des lois de X et Y est la matrice $d \times q$ des termes de covariance entre les composantes des deux vecteurs :

$$\begin{aligned} \mathrm{Cov}(X, Y) &= [\mathrm{Cov}(X_i, Y_j)]_{1 \leq i \leq d,\, 1 \leq j \leq q} \\ &= \mathbb{E}\left[(X - \mathbb{E}X)(Y - \mathbb{E}Y)'\right]. \end{aligned}$$

- La matrice de variance-covariance Σ d'un vecteur aléatoire X est la matrice des covariances croisées entre ses composantes :

$$\Sigma = \mathrm{Cov}(X, X).$$

Loi gaussienne multidimensionnelle

- On souhaite définir une version multidimensionnelle de la loi gaussienne. On ne contente ici de la notion de **loi gaussienne multidimensionnelle** non dégénérée, qui ne nécessite pas de recourir à la notion plus générale de **vecteur gaussien**.
- La fonction caractéristique d'un vecteur aléatoire X de dimension d caractérise la loi de X, elle est définie par

$$\Phi(t) = \mathbb{E}\left(\exp(i\langle t, X\rangle)\right) = \mathbb{E}\left(\exp(it'X)\right)$$

Definition

La fonction caractéristique de la loi normale multidimensionnelle (non dégénérée) de dimension d, d'espérance μ et de matrice de variance-covariance Σ (définie positive) vérifie pour tout $t \in \mathbb{R}^d$,

$$\Phi(\mu, \Sigma)(t) = \exp(it'\mu - \frac{1}{2}t'\Sigma t).$$

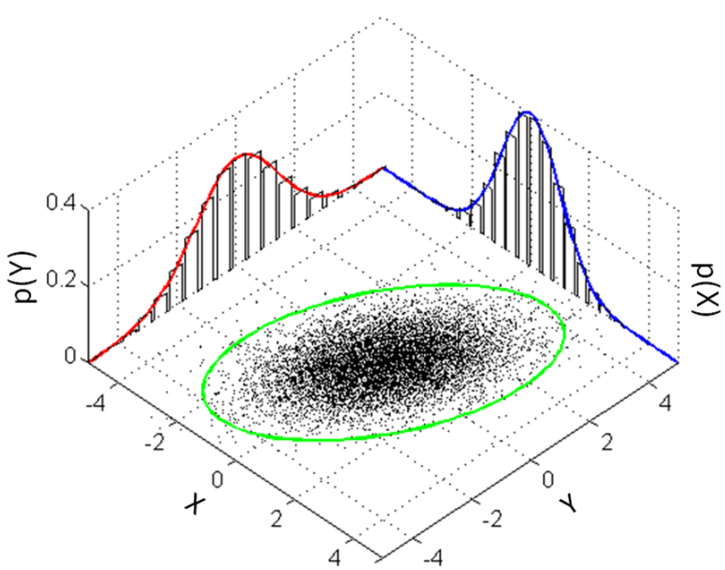
- La loi normale multidimensionnelle est entièrement caractérisée par sa moyenne μ et sa matrice de var-cov Σ. On note $X \sim \mathcal{N}_d(\mu, \Sigma)$.

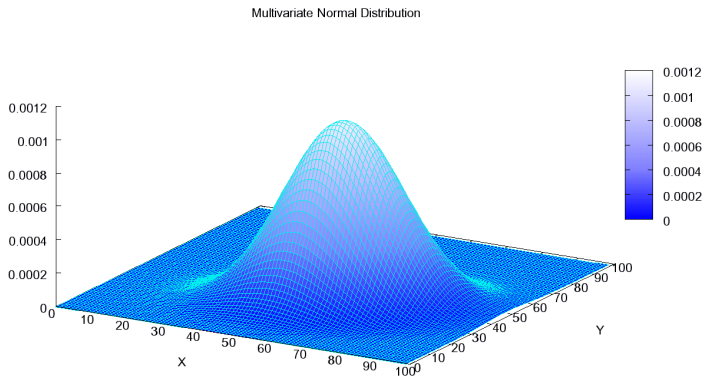
Densité de la loi gaussienne multivariée

Proposition

La distribution de la loi $\mathcal{N}_d(\mu, \Sigma)$ admet une densité f_X pour la mesure de Lebesgue sur $\mathbb{R}^d$ et dans ce cas

$$f_X(x_1, \ldots, x_d) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right].$$





Source : Wikipedia.

Composantes indépendantes

Proposition

Soit $X \sim \mathcal{N}_d(\mu, \Sigma)$, les composantes $X_1, \ldots, X_d$ sont indépendantes si et seulement si Σ est diagonale.

Si les composantes X_j du vecteur X sont indépendantes de loi respectives $\mathcal{N}(\mu_j, \sigma_j^2)$, alors X a une distribution gaussienne multidimentionnelle $\mathcal{N}_d(\mu, \Sigma)$ avec $\mu = (\mu_1, \ldots, \mu_d)$ et $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$.