Optimisation

Thierry Dhorne by Salim Lardjane

1er février 2018

Introduction

Exemple

- problème usuel dans le cours de Statistique
  - ➤ estimer les paramètres d'un modèle
  - ➤ par exemple un modèle de régression linéaire
- le modèle s'écrit généralement

$$Y_r = \beta_0 + \beta_1 x_r + E_r$$

avec

- — Y_r valeurs observables de la variable (aléatoire) expliquée
- — x_r valeurs observées de la variable (aléatoire) explicative
- — E_r valeurs de l'erreurs (aléatoire) résiduels postulés d'espérance nulle (non restrictif) et i.i.d. (assez restrictif)

Moindres carrés

- l'estimation de β_0 et de β_1 peut être réalisée par le maximum de vraisemblance

➤ qui conduit aux critère des moindres carrés

$$S(\beta_0, \beta_1) = \sum_{r=1}^{n} (y_r - \beta_0 - \beta_1 x_r)^2$$

où les y_r désignent des réalisations de la variable aléatoire expliquée

➤ $S(\beta_0, \beta_1)$ quantifie en un certain sens l'ajustement du modèle aux données

- la maximisation de la vraisemblance conduit à la minimisation du critère des moindres carrés

Autre exemple

- estimation des paramètres d'une densité de probabilité
- ➤ à partir d'un échantillon aléatoire issu de la loi correspondante.
- on disposer d'un échantillon observé

$$x_1, x_2, \ldots, x_r, \ldots, x_n$$

➤ issu d'une loi de probabilité de densité $f_\theta(x)$ (ou de masse $p_\theta(k)$)

- prenons l'exemple d'une loi exponentielle

$$f(x) = \lambda e^{-\lambda x} = \frac{1}{\sigma} e^{-\frac{x}{\sigma}}$$

➤ et cherchons à estimer λ (ou σ)

Vraisemblance

- on rappelle que la vraisemblance est vue comme une fonction

— d'argument λ
— de paramètres $\{x_1, x_2, \ldots, x_r, \ldots, x_n\}$

$$\mathcal{L}_{x_1,x_2,\ldots,x_r,\ldots,x_n}(\lambda) = \prod_{r=1}^{n} \lambda e^{-\lambda x_r}$$

- l'estimation du maximum de vraisemblance de λ est la valeur qui maximise

$$\prod_{r=1}^{n} \lambda e^{-\lambda x_r}$$

encore notée l

$$l = \arg\max_{\lambda} \prod_{r=1}^{n} \lambda e^{-\lambda x_r}$$

Problématique

Problème d'optimisation

- le problème considéré porte sur l'optimisation (minimisation ou maximisation) d'une fonction

$$f(x_1, x_2, \ldots, x_m)$$

de m arguments $x_1, x_2, \ldots, x_m$

➤ c'est-à-dire une fonction de $\mathbb{R}^m$ dans $\mathbb{R}$

★ il est équivalent de maximiser f et de minimiser $-f$

➤ on se limite habituellement à l'étude du problème de minimisation.

- dans le cas de l'estimation statistique

➤ les arguments de la fonction à minimiser (- la vraisemblance) sont les paramètres du modèle

➤ les solutions de la minimisation sont les estimations (ou les estimateurs sous forme analytique)

- ces arguments peuvent être contraints ou non

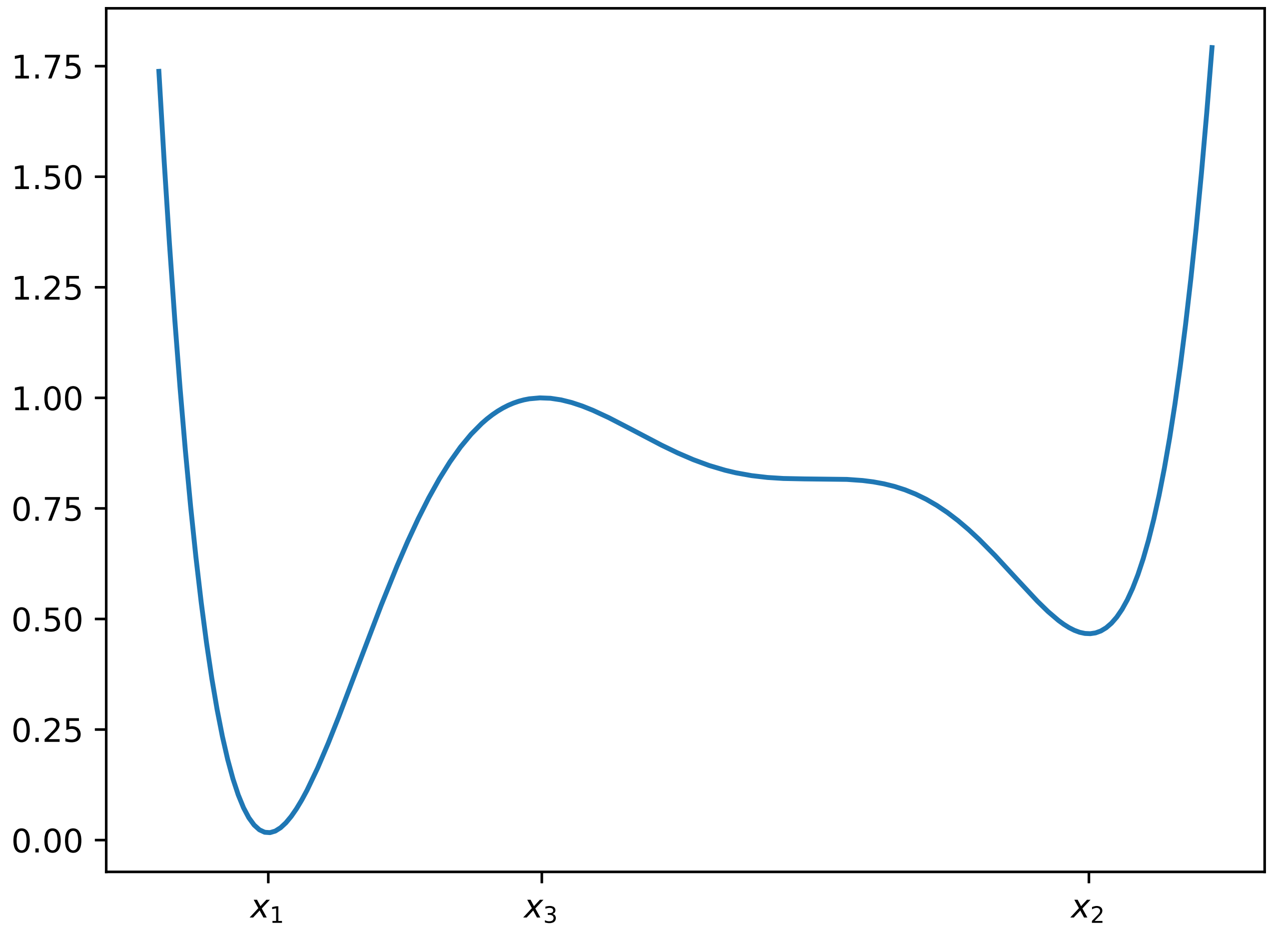
➤ b_0 et b_1 sont a priori non contraints

➤ l est censé être strictement positif

Présentation du problème

- on va présenter le problème dans un cas simple
  - univarié
  - on étendra les intuitions (démonstrations ?) au cas multivarié

Un exemple de fonction à minimiser



- le graphique montre que la fonction possède

— deux minima -> on les appelle x_1 et x_2
— un maximum -> on l'appelle x_3

- ➤ x_1 est un minimum global (absolu)
- ➤ x_2 est un minimum local (relatif)
- ➤ x_3 est un maximum local
- ➤ il n'y a pas (semble-t-il ?) de maximum global

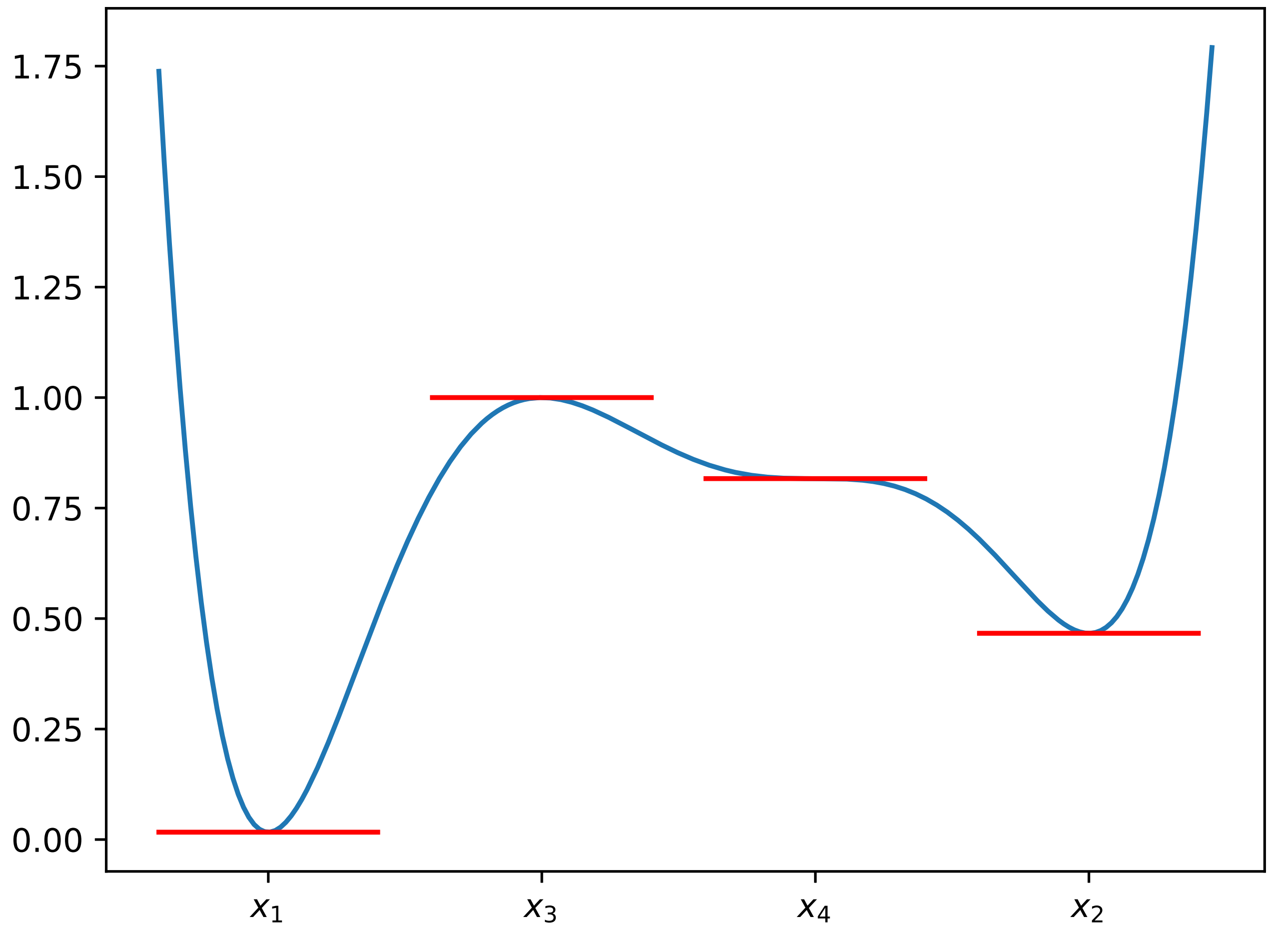
Conditions

- ★ les gens qui ont le sens de l'analyse ont noté que
- ● pour les points concernés les tangentes sont horizontales
- ➤ donc les dérivées sont nulles

$$x^* = \arg\text{ext}(\max ou \min) \Rightarrow \frac{df}{dx}(x^*) = 0$$

- ★ mais la réciproque est fausse

Points de dérivée nulle



Points selles (ou d'inflexion) - Constatations

- on retrouve sur le graphique les trois points prévus
  - ➤ plus un quatrième dont on peut démontrer que la dérivée s'annule pour lui
  - ➤ on l'appelle x_4
- on s'aperçoit que pour ce point (à vélo) la vitesse diminue puis réaugmente
  - ➤ la dérivée première va jusqu'à s'annuler mais reprend des valeurs positives
  - ➤ elle passe donc par un minimum
  - ★ elle s'annule ponctuellement sans changer de signe
- à la différence des minima pour lesquels
- la vitesse diminue et continue de diminuer
  - ➤ elle ne passe donc pas par un minimum
  - ★ elle s'annule en changeant de signe

Conclusions

- on a vu que

$$x^* = \arg \mathrm{ext}(\max ou \min) \Rightarrow \frac{df}{dx}(x^*) = 0$$

➢ mais la réciproque n'est pas vraie

$$\frac{df}{dx}(x^*) = 0 \not\Rightarrow x^* = \arg \mathrm{ext}(\max ou \min)$$

- pour les points d'inflexion

➢ on a vu que la dérivée s'annule mais ne change pas de signe

★ elle passe donc (elle) par un extremum

- et pour les extrema

➢ elle s'annule et change de signe sans atteindre d'extremum

Conditions d'extremum

- pour un extremum la dérivée change de signe (sans passer par un extremum) et donc sa dérivée vérifie

$$\frac{d^2 f}{dx^2}(x^*) > 0$$

ou

$$\frac{d^2 f}{dx^2}(x^*) < 0$$

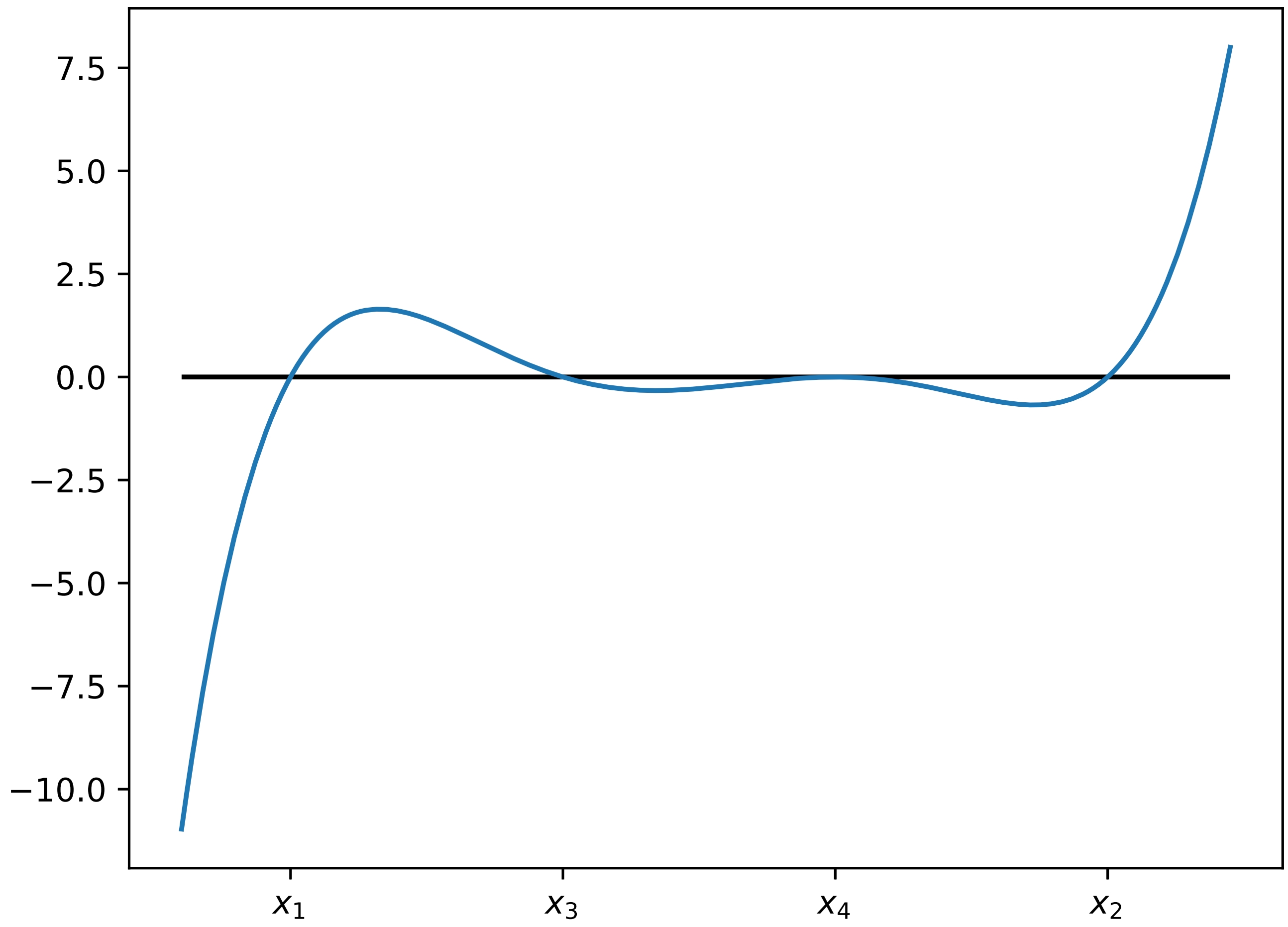
➤ à ajouter à la contrainte sur la dérivée première

$$\frac{df}{dx}(x^*) = 0$$

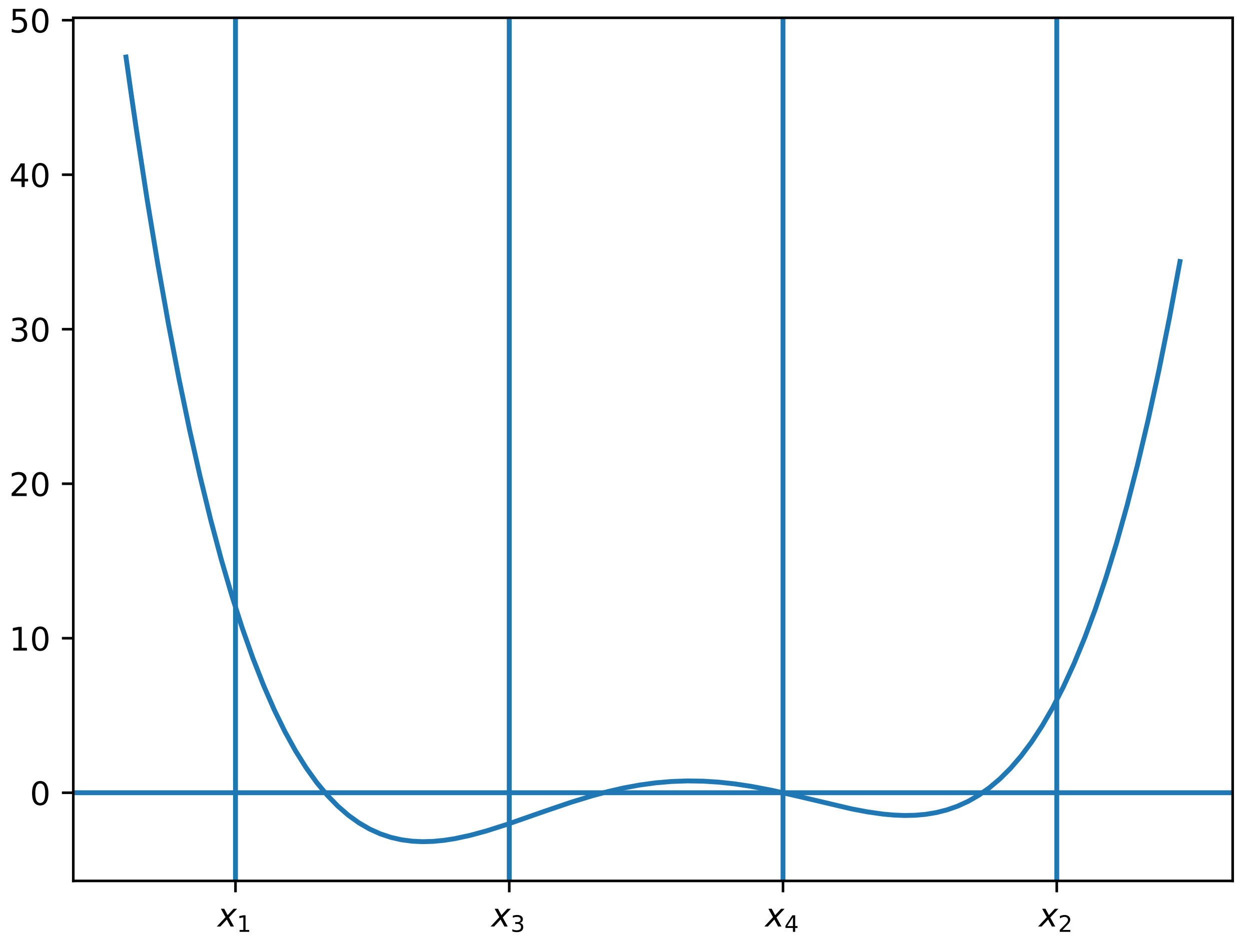
- à l'inverse pour les points d'inflexion, pour la valeur précise d'annulation de la dérivée première on a

$$\frac{d^2 f}{dx^2}(x^*) = 0$$

Vérification avec la dérivée première



Vérification avec la dérivée seconde



- on peut étendre le raisonnement à la minimisation d'une fonction de plusieurs variables

$$f(x_1, x_2, \ldots, x_m)$$

➤ une condition nécessaire pour qu'un point $x^* = (x_1^*, x_2^*, \ldots, x_m^*)$ corresponde à un minimum est

$$\frac{\partial f}{\partial x_1}(x^*) = \frac{\partial f}{\partial x_2}(x^*) = \ldots = \frac{\partial f}{\partial x_m}(x^*) = 0$$

★ les solutions de cette équation peuvent également correspondre à des maxima ou a des points-selles

➤ il s'agit donc comme plus haut d'une condition nécessaire

Condition suffisante

- une condition suffisante pour qu'une solution x^* de l'équation précédente corresponde à un minimum

➤ est que la matrice $H(x^*)$ d'éléments

$$h_{ij}(x^*) = \frac{\partial^2 f}{\partial x_i \partial x_j}(x^*)$$

soit définie positive

- cette matrice de dimension m est naturellement symétrique

★ elle est appelée matrice hessienne (ou hessien) de f

Mais savez-vous bien ce qu'est une matrice définie positive ?

Applications

Estimation par ML du paramètre d'une loi exponentielle

- la vraisemblance de l'échantillon a été indiquée plus haut

$$\mathcal{L}_{x_1,x_2,\ldots,x_r,\ldots,x_n}(\lambda) = \prod_{r=1}^{n} \lambda e^{-\lambda x_r}$$

➤ on cherche la valeur l qui maximise en λ l'expression précédente

➤ on va d'abord prendre le ln de cette expression

★ en effet la fonction ln

- ne modifie pas l'argument du maximum car c'est une fonction strictement monotone (croissante)
- simplifie les calculs en « supprimant » les exponentiations

- l'expression devient

$$\begin{aligned} \ln(\mathscr{L})_{x_1,x_2,\ldots,x_r,\ldots,x_n}(\lambda) &= \sum_{r=1}^{n}(\ln(\lambda) - \lambda x_r) \\ &= n(\ln(\lambda) - \lambda\sum_{r=1}^{n} x_r \end{aligned}$$

- ➤ on dérive cette expression par rapport à λ

$$\frac{\partial \ln(\mathscr{L})}{\partial \lambda} = n\frac{1}{\lambda} - \sum_{r=1}^{n} x_r$$

- ➤ on annule cette dérivée

$$n\frac{1}{l} - \sum_{r=1}^{n} x_r = 0$$

Solution

- la solution est

$$l = \frac{n}{\sum_{r=1}^{n} x_r} = \frac{1}{\bar{x}}$$

➤ la dérivée seconde de la ln −vraisemblance est

$$\begin{aligned} \frac{\partial^2 \ln(\mathscr{L})}{\partial \lambda^2} &= -\frac{n}{l^2} \\ &< 0 \end{aligned}$$

- la dérivée seconde est toujours négative

➤ la vraisemblance atteint donc bien un maximum au point concerné

Régression linéaire simple

- on veut minimiser la somme des carrés des écarts

$$S(\beta_0, \beta_1) = \sum_{r=1}^{n} (y_r - \beta_0 - \beta_1 x_r)^2$$

➤ la condition nécessaire de minimum s'écrit

$$\begin{cases} \frac{\partial S}{\partial \beta_0}(b_0, b_1) & = & 0 \\ \frac{\partial S}{\partial \beta_1}(b_0, b_1) & = & 0 \end{cases}$$

➤ soit

$$\begin{cases} \sum_{r=1}^{n} (y_r - b_0 - b_1 x_r) & = & 0 \\ \sum_{r=1}^{n} x_r (y_r - b_0 - b_1 x_r) & = & 0 \end{cases}$$

Matrice hessienne

- la matrice hessienne est

$$2\begin{pmatrix} \sum_{r=1}^{n} 1 & \sum_{r=1}^{n} x_r \\ \sum_{r=1}^{n} x_r & \sum_{r=1}^{n} x_r^2 \end{pmatrix}$$

➤ les valeurs propres de la matrice vérifient l'équation

$$\begin{aligned} (n-\alpha)(\sum_{r=1}^{n} x_r^2 - \alpha) - (\sum_{r=1}^{n} x_r)^2 &= 0 \\ \alpha^2 - (\sum_{r=1}^{n} x_r^2 + n)\alpha + n\sum_{r=1}^{n} x_r^2 - (\sum_{r=1}^{n} x_r)^2 &= 0 \end{aligned}$$

➤ elles sont égales à

$$n + \sum_{r=1}^{n} x_r^2 \pm \sqrt{(n - \sum_{r=1}^{n} x_r^2) + 4(\sum_{r=1}^{n} x_r)^2}$$

Signe des valeurs propres - 1

- la première valeur propre

$$n + \sum_{r=1}^{n} x_r^2 + \sqrt{(n - \sum_{r=1}^{n} x_r^2) + 4(\sum_{r=1}^{n} x_r)^2}$$

  ➢ est manifestement positive

- montrons que la seconde l'est aussi c'est-à-dire que

$$n + \sum_{r=1}^{n} x_r^2 > \sqrt{(n - \sum_{r=1}^{n} x_r^2) + 4(\sum_{r=1}^{n} x_r)^2}$$

  ➢ soit en élevant au carré

$$(n + \sum_{r=1}^{n} x_r^2)^2 > (n - \sum_{r=1}^{n} x_r^2) + 4(\sum_{r=1}^{n} x_r)^2$$

Signe des valeurs propres - 2

- ce qui est équivalent à

$$
\begin{aligned}
n^2+(\sum_{r=1}^{n} x_r^2)^2+2n\sum_{r=1}^{n} x_r^2 &> n^2+(\sum_{r=1}^{n} x_r^2)^2-2n\sum_{r=1}^{n} x_r^2+4(\sum_{r=1}^{n} x_r)^2 \\
&\Updownarrow \\
4n\sum_{r=1}^{n} x_r^2 &> 4(\sum_{r=1}^{n} x_r)^2 \\
&\Updownarrow \\
n\sum_{r=1}^{n} x_r^2-(\sum_{r=1}^{n} x_r)^2 &> 0 \\
&\Updownarrow \\
\sum_{r=1}^{n} (x_r-\bar{x})^2 &> 0
\end{aligned}
$$

★ cette dernière égalité est évidemment vérifiée et donc la matrice est s.d.p.

Conclusions

- dans les deux cas précédents, les équations spécifiant l'optimum ont pu être résolues explicitement.
- ce n'est pas toujours le cas en pratique
- ➤ comme pour les paramètres de la loi Γ
- ★ d'où l'intérêt des algorithmes de résolution numérique