Optimisation à finalité statistique

Salim Lardjane

Université de Bretagne-Sud

Cours 1 - Généralités

Introduction

Un problème considéré de façon usuelle dans les cours de Statistique consiste à déterminer les jeux de paramètres dans un modèle de régression linéaire.

Le modèle est généralement mis sous la forme

$$y_i = \alpha + \beta t_i + \varepsilon_i$$

où les y_i sont les observations de la variable expliquée, t_i les valeurs de la variable explicative et ε_i les termes d'erreur, supposés i.i.d. et d'espérance nulle. Les t_i sont supposés déterministes (fixes).

Introduction

La détermination de α et β peut être abordée de plusieurs façons.

La plus commune est de définir un *critère d'ajustement* qui quantifie en un certain sens l'ajustement du modèle aux données.

Introduction

Un critère usuel est le *critère des moindres carrés*

$$s(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta t_i)^2$$

où les y_i désignent ici des réalisations de la variable expliquée.

Les valeurs plus petites de $s(\alpha, \beta)$ indiquent un meilleur ajustement aux données.

Par conséquent, il est raisonnable d'adopter pour estimation de (α, β) la valeur qui minomise $s(\alpha, \beta)$, qu'on peut noter

$$\widehat{(\alpha, \beta)} = \arg \min_{(\alpha,\beta)} s(\alpha, \beta)$$

On retrouve là la procédure des *moindres carrés*.

Introduction

Un autre type de problème commun en Statistique est l'estimation des paramètres d'une densité de probabilité à partir d'un échantillon aléatoire issu de la loi correspondante.

Introduction

On peut par exemple disposer d'un échantillon $x_1, \ldots, x_n$ issu d'une loi de probabilité exponentielle, donc de densité

$$f(x) = \lambda e^{-\lambda x} \mathbb{I}_{]0,+\infty[}(x)$$

et souhaiter déterminer λ.

Pour ce faire, on considère que $x_1, \ldots, x_n$ sont des réalisation d'un échantillon i.i.d. d'observations et on forme la *vraisemblance*

$$\mathcal{L}(x_1, \ldots, x_n; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

On adopte alors pour estimation de λ la valeur

$$\widehat{\lambda} = \arg\max_{\lambda} \mathcal{L}(x_1, \ldots, x_n; \lambda)$$

Cette méthode est appelée méthode du maximum de vraisemblance et est à l'origine de nombreux problèmes d'optimisation rencontrés en Statistique.

Problème d'optimisation

Sous sa forme la plus générale, le problème qu'on considère porte sur l'*optimisation* (minimisation ou maximisation) d'une fonction $f(\theta_1, \ldots, \theta_m)$ de m *paramètres* $\theta_1, \ldots, \theta_m$, c'est-à-dire de

$$f : \mathbb{R}^m \to \mathbb{R}$$

Notons d'abord qu'il est équivalent de maximiser f et de minimiser $-f$. On peut donc se limiter à l'étude du problème de *minimisation*.

Problème d'optimisation

Notons ensuite que les valeur prises par les paramètres peuvent être *contraintes* ou *non contraintes*.

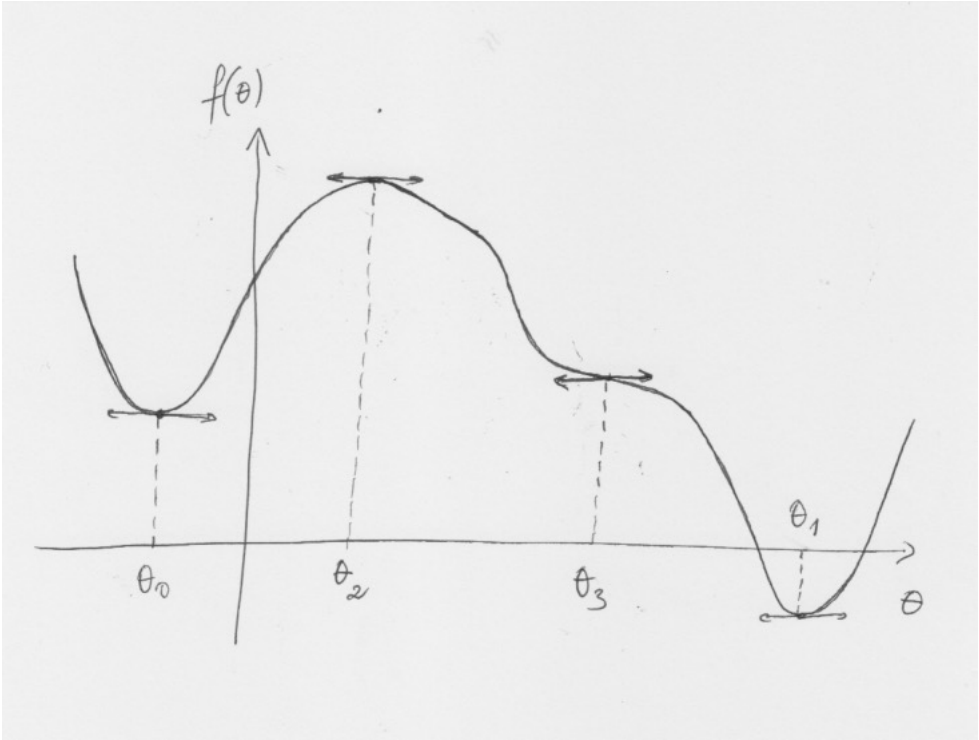
Par exemple, dans le modèle de régression linéaire vu plus haut, les paramètres ne sont pas contraints.

Par contre, dans l'exemple d'estimation par MV, λ est contraint à être strictement positif.

Problème d'optimisation

On va introduire les concepts auxquels nous auront recours par la suite sur l'exemple d'une fonction *d'un seul paramètre*.

Plus précisément, sur



Problème d'optimisation

Ce graphique montre que la fonction possède

– *deux minima* (en θ_0 et θ_1)

– un *maximum* en θ_2

– un *point d'inflexion* (point-selle) en θ_3.

θ_0 est un *minimum local* car $f(\theta_0)$ est inférieur à $f(\theta)$ pour tout θ dans un voisinage de θ_0.

θ_1 est un *minimum global* car $f(\theta_1) \leq f(\theta)$ pour tout θ.

Une question majeure dans les problèmes d'optimisation complexes consiste à *déterminer si l'on a identifié un minimum local ou global*.

Problème d'optimisation

L'approche classique pour déterminer les valeurs de θ_0 et de θ_1 est de remarquer que le *gradient* (la dérivée) de f est nul en θ_0 et θ_1.

Ainsi, θ_0 et θ_1 sont solutions de l'équation au premier ordre

$$\frac{df}{d\theta}(\theta) = 0$$

Comme on peut le voir sur la figure, la valeur θ_2 qui correspond à un maximum local et θ_3 qui correspond à un point d'inflexion vérifient également cette équation.

Problème d'optimisation

Ainsi, la condition précédente est une condition *nécessaire* mais *non suffisante* pour qu'un point corresponde à un minimum.

Toutefois, en examinant la figure, on voit qu'en θ_0 et θ_1, le gradient change de signe du négatif au positif ; en θ_2, il change de signe du positif au négatif et en θ_3, il ne change pas de signe.

Ainsi, au voisinage d'un minimum, le gradient est une fonction *croissante*.

Problème d'optimisation

Or le taux de variation du gradient est mesuré par la dérivée seconde.

Par conséquent, pour qu'un point θ^* solution de l'équation au premier ordre précédente corresponde à un minimum, on doit avoir

$$\frac{d^2 f}{d\theta^2}(\theta^*) > 0.$$

Problème d'optimisation

Ces idées peuvent être étendues à la minimisation d'une fonction de plusieurs paramètres $f(\theta_1, \ldots, \theta_m)$.

Une condition nécessaire pour qu'un point $\theta^* = (\theta_1^*, \ldots, \theta_m^*)'$ corresponde à un minimum est que

$$\frac{\partial f}{\partial \theta_1}(\theta^*) = \frac{\partial f}{\partial \theta_2}(\theta^*) = \cdots = \frac{\partial f}{\partial \theta_m}(\theta^*) = 0$$

Les solutions de cette équation peuvent également correspondre à des maxima ou a des points-selles.

Problème d'optimisation

Une condition *suffisante* pour qu'une solution θ^* de l'équation précédente corresponde à un minimum est que la matrice $H(\theta^*)$ d'éléments

$$h_{ij}(\theta^*) = \frac{\partial^2 f}{\partial\theta_i \partial\theta_j}(\theta^*)$$

soit *définie positive*.

Cette condition généralise celle du cas univarié. $H(\theta^*)$ est appelée *matrice hessienne* de f en θ^*. Elle est symétrique et de dimensions $m \times m$.

Quelques exemples

Considérons d'abord le problème d'estimation par maximum de vraisemblance du paramètre d'une loi exponentielle.

La fonction de vraisemblance s'écrit

$$\mathcal{L}(x_1, \ldots, x_n; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

On souhaite estimer λ par la valeur $\widehat{\lambda}$ qui maximise $\mathcal{L}(\lambda)$ ou, de façon équivalente, qui minimise $-\log \mathcal{L}(\lambda)$.

On note parfois $\log \mathcal{L}(\lambda) = \ell(\lambda)$.

On souhaite donc minimiser

$$s(\lambda) = -\log \mathcal{L}(\lambda) = -\ell(\lambda) = \lambda \sum_{i=1}^{n} x_i - n \log(\lambda)$$

Quelques exemples

Une condition nécessaire pour que $\widehat{\lambda}$ corresponde à un minimum de $s(\lambda)$ est que

$$\frac{ds}{d\lambda}(\widehat{\lambda}) = 0$$

ce qui s'écrit

$$\sum_{i=1}^{n} x_i - \frac{n}{\widehat{\lambda}} = 0$$

d'où

$$\widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} xi} = \frac{1}{\overline{x}_n}$$

ce qui correspond bien à un minimum de $s(\lambda)$ puisque

$$\frac{d^2s}{d\lambda^2}(\widehat{\lambda}) = \frac{n}{\widehat{\lambda}^2} > 0.$$

Quelques exemples

Reprenons à présent le problème d'estimation de (α, β) dans le modèle de régression linéaire simple.

On souhaite estimer (α, β) par $(\widehat{\alpha}, \widehat{\beta})$ qui minimise

$$s(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta t_i)^2$$

où y_i désignent les réalisations de la variable expliquée.

La condition nécessaire de minimum s'écrit

$$\left\{ \begin{array}{rcl} \frac{\partial s}{\partial \alpha}(\widehat{\alpha}, \widehat{\beta}) & = & 0 \\ \frac{\partial s}{\partial \beta}(\widehat{\alpha}, \widehat{\beta}) & = & 0 \end{array} \right.$$

Quelques exemples

Autrement dit

$$\left\{ \begin{array}{rcl} -2\sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta}t_i) & = & 0 \\ -2\sum_{i=1}^{n} t_i(y_i - \widehat{\alpha} - \widehat{\beta}t_i) & = & 0 \end{array} \right.$$

d'où on déduit après un calcul simple

$$\widehat{\alpha} = \bar{y}_n - \widehat{\beta}\bar{t}_n$$

et

$$\widehat{\beta} = \frac{C(t,y)}{C(t,t)}$$

où

$$\begin{aligned} C(t,y) &= \sum_{i=1}^{n} t_i y_i - \sum_{i=1}^{n} t_i \sum_{i=1}^{n} y_i/n \\ C(t,t) &= \sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i\right)^2 /n \end{aligned}$$

Quelques exemples

La *matrice hessienne* de s en $(\widehat{\alpha}, \widehat{\beta})$ s'écrit

$$\begin{pmatrix} 2n & 2\sum_{i=1}^{n} t_i \\ 2\sum_{i=1}^{n} t_i & 2\sum_{i=1}^{n} t_i^2 \end{pmatrix}$$

On peut montrer que $H(\widehat{\alpha}, \widehat{\beta})$ est *définie positive* (par exemple, à l'aide du *critère de Sylvester*) et par conséquent que $(\widehat{\alpha}, \widehat{\beta})$ correspond à un *minimum* de $s(\alpha, \beta)$.

Quelques exemples

Dans les deux cas précédents, les équations spécifiant le minimum ont pu être *résolues explicitement*.

Ce n'est cependant pas toujours le cas en pratique.

Pour de nombreux problèmes, on doit avoir recours à d'autres approches que le calcul direct pour déterminer des estimations. On parle alors de *procédures ou algorithmes de minimisation*.