Analyse de données pour les systèmes embarqués
Recueil de formules

Lionel Lacassagne

LIP6
Université Pierre et Marie Curie (Paris 6)

Plan

Observations :

- La mesure de grandeurs physiques (position, vitesse, accélération, distance par rapport à un plan, pression, luminosité) est entachée d'erreur
- il faut les traiter (filtrer et les analyser) pour pouvoir s'en servir correctement.

Quatre aspects :

- analyse statistique (moyenne, médian, écart-type, min/max)
- régression (linéaire, quadratique, cubique)
- interpolation : calcul d'une valeur dans un ensemble (linéaire, cubique)
- extrapolation : calcul d'une valeur en dehors d'un ensemble (ex : position future)

Objectifs :

- savoir quand utiliser une méthode (connaitre ce qu'elle apporte à la mesure d'une grandeur)
- connaitre la complexité des différentes méthodes pour faire des compromis

Analyse statistique : moyenne, variance et écart type

La moyenne $\overline{x}$ d'un ensemble de N valeurs et la variance σ_x sont par définition

$$\overline{x} \triangleq \frac{1}{N}\sum_i x_i \quad \text{et} \quad \sigma_x^2 \triangleq \frac{1}{N}\sum_i (x_i - \overline{x})^2$$

Le calcul de la variance nécessite de balayer deux fois les données. Une première fois pour calculer la moyenne et une seconde pour calculer la différence entre chaque point et la moyenne. Deux problèmes :

- le calcul de la variance nécessite de stocker toutes les valeurs nécessaires dans un tableau. Pas de calcul *à la volée* (*on the fly*)
- Comme la vitesse de traitement est souvent limitée par le nombre d'accès mémoire (*memory bound*), 2 balayages prennent 2 fois plus de temps qu'un seul.

Reformulation sous forme de sommes :

$$S = \sum_{i=1}^{N} 1 = N, \qquad S_x = \sum_{i=1}^{N} x_i, \qquad S_{xx} = \sum_{i=1}^{N} x_i^2$$

Les formules de la moyenne et de la variance deviennent

$$\overline{x} = \frac{Sx}{S} \qquad \sigma_x^2 = \frac{SS_{xx} - S_x^2}{S^2}$$

Régression linéaire #1

On veut calculer l'équation de la droite $y = a + bx$ passant au mieux (au sens des moindres carrée) par un nuage de points. Fonction de mérite χ^2

$$\chi^2(a,b) = \sum_{i=1}^{N} \left(\frac{y_i - a - bx_i}{\sigma_i} \right)^2$$

On veut minimiser $\chi^2 \Leftrightarrow$ les dérivées s'annulent

$$\frac{\partial \chi^2}{\partial a} = 0 \quad \Rightarrow \quad -2 \sum_{i=1}^{N} \frac{y_i - a - bx_i}{\sigma_i^2} = 0$$

$$\frac{\partial \chi^2}{\partial b} = 0 \quad \Rightarrow \quad -2 \sum_{i=1}^{N} \frac{x_i(y_i - a - bx_i)}{\sigma_i^2} = 0$$

Régression linéaire #2

Cela revient à résoudre le système linéaire 2×2 :

$$\begin{aligned} aS + bS_x &= S_y \\ aS_x + bS_{xx} &= S_{xy} \end{aligned}$$

En notation matricielle :

$$\begin{bmatrix} S & S_x \\ S_x & S_{xx} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} S_y \\ S_{xy} \end{bmatrix}$$

Le système est inversible si $\Delta \neq 0$:

$$\begin{aligned} \Delta &\equiv SS_{xx} - S_x^2 \\ a &= \frac{S_{xx}S_y - S_xS_{xy}}{\Delta} \\ b &= \frac{SS_{xy} - S_xS_y}{\Delta} \end{aligned}$$

Régression linéaire #3

Autres grandeurs statistiques associés :

- variance de a et b

$$\begin{aligned}\sigma_a^2 &= S_{xx}/\Delta \\ \sigma_b^2 &= S/\Delta\end{aligned}$$

- covariance de a et b :

$$\mathrm{cov}(a, b) = -S_x/\Delta$$

- coefficient de corrélation (entre -1 et 1)

$$r_{ab} = \frac{-S_x}{\sqrt{SS_{xx}}}$$

Remarque : plus il y a de points (pouvant être issu de la droite), meilleur est le *fit* (r_{ab} tend vers 1)

Régression non linéaire

Quatre modèles possibles :

- regression linéaire : $y = a + bx$
- courbe exponentielle : $y = ae^{bx}$
- courbe logarithmique : $y = a + b\ln(x)$
- courbe puissance : $y = ax^b$

Il suffit de modifier ce que l'on somme :

Regression	A	X_i	Y_i
linéaire	a	x_i	y_i
exponentielle	$\ln a$	x_i	$\ln y_i$
logarithmique	a	$\ln x_i$	y_i
puissance	$\ln a$	$\ln x_i$	$\ln y_i$

Régressions multiples

Deux variables indépendantes :

$$t = a + bx + cy \tag{1}$$

$$\begin{bmatrix} S & S_x & S_y \\ S_x & S_{xx} & S_{xy} \\ S_y & S_{xy} & S_{yy} \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} S_t \\ S_{xt} \\ S_{yt} \end{bmatrix}$$

Trois variable indépendantes :

$$t = a + bx + cy + dz$$

$$\begin{bmatrix} S & S_x & S_y & S_z \\ S_x & S_{xx} & S_{xy} & S_{xz} \\ S_y & S_{xy} & S_{yy} & S_{yz} \\ S_z & S_{xz} & S_{yz} & S_{zz} \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} S_t \\ S_{xt} \\ S_{yt} \\ S_{zt} \end{bmatrix}$$

Régressions polynomiales #1

Régression parabolique / quadratique

$$t = a + bx + cx^2$$

$$\begin{bmatrix} S & S_x & S_{x^2} \\ S_x & S_{x^2} & S_{x^3} \\ S_{x^2} & S_{x^3} & S_{x^4} \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} S_t \\ S_{xt} \\ S_{x^2 t} \end{bmatrix}$$

$$R^2 = \frac{aS_t + bS_{xt} + cS_{x^2 t} - \frac{1}{n} S_t^2}{S_{t^2} - \frac{1}{n} S_t^2}$$

Régressions polynomiales #2

Régression cubique

$$t = a + bx + cx^2 + dx^3$$

$$\begin{bmatrix} S & S_x & S_{x^2} & S_{x^3} \\ S_x & S_{x^2} & S_{x^3} & S_{x^4} \\ S_{x^2} & S_{x^3} & S_{x^4} & S_{x^5} \\ S_{x^3} & S_{x^4} & S_{x^5} & S_{x^6} \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} S_t \\ S_{xt} \\ S_{x^2t} \\ S_{x^3t} \end{bmatrix}$$

$$R^2 = \frac{aS_t + bS_{xt} + cS_{x^2t} + dS_{x^3t} - \frac{1}{n}S_t^2}{S_{t^2} - \frac{1}{n}S_t^2}$$

Attention plus le degré est élevé et plus les variations du polynôme seront grandes (lapalissade) $\Rightarrow$ plus il faut d'échantillons

Interpolation linéaire générale (espace 2D)

- Soient deux points (x_0, y_0) and (x_1, y_1) que l'on veut interpoler linéairement.
- Pour une valeur x dans l'intervalle $[x_0, x_1]$, la valeur y le long de la ligne droite est donnée par :

$$\frac{y - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0}$$

soit

$$y = y_0 + (y_1 - y_0)\frac{x - x_0}{x_1 - x_0}$$

La formule peut être interprétée comme une somme pondérée :

$$y = (1 - \alpha) \times y_0 + \alpha \times y1 \quad \text{avec} \quad \alpha = \frac{x - x_0}{x_1 - x_0}$$

La formule est valide en dehors de l'intervalle.

Interpolation linéaire sur une grille (espace 1D)

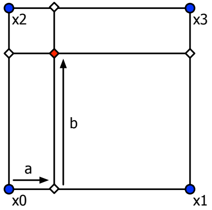
- Souvent les points sont équi-espacés (grille).
- Il y a un changement de notation : le point n a la valeur x_n et le point $n+1$ a la valeur x_{n+1}.
- La formule d'interpolation devient :

$$x = (1-\alpha) \times x_n + \alpha \times x_{n+1} \quad \text{avec} \quad \alpha \in [0,1]$$

- on vérifie que si $\alpha = 0, x = x_n$ et si $\alpha = 1, x = x_{n+1}$

Interpolation bi-linéaire sur une grille 2D

l'interpolation bi-linéaire est séparable



calcul horizontal puis vertical

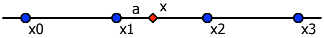
$$\begin{aligned} x_{01} &= (1-\alpha)x_0 + \alpha x_1 \\ x_{23} &= (1-\alpha)x_2 + \alpha x_3 \\ x &= (1-\beta)x_{01} + \beta x_{23} \end{aligned}$$

calcul vertical puis horizontal

$$\begin{aligned} x_{02} &= (1-\beta)x_0 + \beta x_2 \\ x_{13} &= (1-\beta)x_1 + \beta x_3 \\ x &= (1-\alpha)x_{02} + \alpha x_{13} \end{aligned}$$

Interpolation cubique (1D)

Soient x_1, x_2, x_3, x_4 quatre points sur une grille de pas unitaire. On confond par soucis de simplification le nom et la valeur des points.



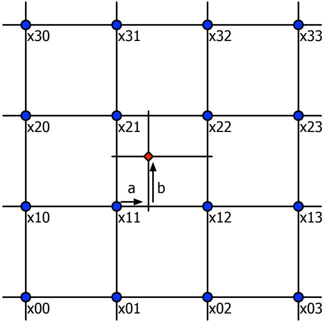
Soient les quatre coefficients

$$\begin{aligned}
c_1(\alpha) &= -\frac{1}{2}(\alpha^3 - 2\alpha^2 + \alpha) \\
c_2(\alpha) &= +\frac{1}{2}(3\alpha^3 - 5\alpha^2 + 2) \\
c_3(\alpha) &= -\frac{1}{2}(3\alpha^3 - 4\alpha^2 - \alpha) \\
c_4(\alpha) &= +\frac{1}{2}(\alpha^3 - \alpha^2)
\end{aligned}$$

Alors l'interpolation cubique en x décalé d'une quantité α par rapport à x1 est :

$$x = c_1 x_1 + c_2 x_2 + c_3 x_3 + c_4 x_4$$

Interpolation bi-cubique (2D)



- l'interpolation bi-cubique est aussi séparable
- elle est très utilisée en retouche photo et en traitement d'images, lors qu'il faut faire un *resize* d'une image

Gradient interpolé cubique

Le gradient s'obtient en dérivant les formules de coefficients puis en appliquant le même motif de calcul.

$$
\begin{aligned}
g_1(\alpha) &= -\frac{1}{2}(3\alpha^2 - 4\alpha + 1) \\
g_2(\alpha) &= +\frac{1}{2}(9\alpha^2 - 10\alpha) \\
g_3(\alpha) &= -\frac{1}{2}(9\alpha^2 - 8\alpha - 1) \\
g_4(\alpha) &= +\frac{1}{2}(3\alpha^2 - 2\alpha)
\end{aligned}
$$

Gradient interpolé bi-cubique (2D)

Il faut calculer la dérivée dans une direction et le lissage dans l'autre :

- gradient en x : gradient interpolé en x et lissage en y
- gradient en y : gradient interpolé en y et lissage en x

On ne dérive que dans une direction : dans l'autre on lisse.

Résolution de systèmes linéaires

De la forme $Ax = b$, avec A une matrice carrée, x et b deux vecteurs (b est appelé *second membre*).

- Deux cas particuliers : 2×2 et 3×3
- cas général (> 3)

Cas 2×2 :

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

$$\Delta = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \quad \Delta_1 = \begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix} \quad \Delta_2 = \begin{vmatrix} a_{11} & b_{12} \\ a_{21} & b_{22} \end{vmatrix}$$

Si $\Delta \neq 0$ alors :

$$x_1 = \frac{\Delta_1}{\Delta} \quad x_2 = \frac{\Delta_2}{\Delta}$$

Résolution de systèmes linéaires 3×3

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \quad \Delta = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

$$\Delta_1 = \begin{vmatrix} b_{11} & a_{12} & a_{13} \\ b_{21} & a_{22} & a_{23} \\ b_{31} & a_{32} & a_{33} \end{vmatrix} \quad \Delta_2 = \begin{vmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{vmatrix} \Delta_3 = \begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{vmatrix}$$

Règle de Sarrus pour développer les déterminants 3×3. Développement par rapport à la première ligne : faire le produit des termes sur une diagonale. Sommer positivement si la diagonale va vers la droite, négativement si la diagonale va vers la gauche.

$$\Delta = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{12}a_{22}a_{31}$$

Si $\Delta \neq 0$ alors :

$$x_1 = \frac{\Delta_1}{\Delta} \quad x_2 = \frac{\Delta_2}{\Delta} \quad x_3 = \frac{\Delta_3}{\Delta}$$

Résolution de systèmes linéaires

- Cas général : algorithme du pivot de Gauss
- cas particulier #1 : décomposition LU (pas de contrainte sur les matrices). Pivotage nécessaire (choix d'un pivot pour maximiser la précision des calculs).
- cas particulier #2 : décomposition de Choleski (pour les matrices symétrique semi-définie positives). Pas besoin de pivoter et Deux fois moins de calculs.

- Les méthodes LU et Choleski sont très utiles lorsque pour une même matrice, il faut résoudre le système pour plusieurs second-membres.
- les systèmes à résoudre ici sont tous symétriques $\Rightarrow$ Choleski

Décomposition de Choleski

- Trois étapes : factorisation : décomposition en 2 matrices triangulaires
- descente : résolution d'un premier système linéaire intermédiaire
- remontée : résolution du système linéaire

Choleski - formules

On cherche L tel que $A = LL^t$ avec

$$L = \begin{bmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix}$$

Pour $i = 1$, on détermine la première colonne de L :

$$l_{11} = \sqrt{a_{11}} \quad \text{et} \quad l_{j1} = \frac{a_{ij}}{l_{11}} \text{pour } j = 2:n$$

Pour $i = 2:n$, on procède colonne par colonne :

$$l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2} \quad \text{et} \quad l_{ji} = \frac{a_{ij} - \sum_{k=1}^{i-1} l_{ik} l_{kj}}{l_{ii}}$$

Choleski - algorithme

Factorisation de $A = LL^t$

Algorithm 1: Choleski - factorisation

for $i = 1 : n$ **do**
- $s \leftarrow 0$
- **for** $k = 1 : i-1$ **do**
  - $s \leftarrow s + L(i,k)^2$
- $L(i,i) = \sqrt{A(i,i) - s}$
- **for** $j = i+1 : n$ **do**
  - $s \leftarrow 0$
  - **for** $k = 1 : i-1$ **do**
    - $s \leftarrow s + L(i,k) \times L(j,k)$
  - $L(j,i) \leftarrow (A(i,j) - s)/L(i,i)$

Choleski - descente

Résolution du sytème triangulaire (inférieur) $Ly = b$

Algorithm 2: Choleski - descente

---

for $i = 1 : n$ **do**
  $s \leftarrow b(i)$
  for $j = 1 : i - 1$ **do**
    $s \leftarrow s - L(i, j) \times b(j)$
  $b(i) \leftarrow s/L(i, i)$

Résolution du sytème triangulaire (supérieur) $Lx = y$

Algorithm 3: Choleski - remonté

---

for $i = n : 1 : -1$ **do**
  $s \leftarrow b(i)$
  for $j = n : i + 1 : -1$ **do**
    $s \leftarrow s - L(i, j) \times b(j)$
  $b(i) \leftarrow s/L(i, i)$

Exemple de factorisation de Choleski 4×4

La matrice symétrique A

$$A = \begin{pmatrix} 1 & 2 & 4 & 7 \\ 2 & 13 & 23 & 38 \\ 4 & 23 & 77 & 122 \\ 7 & 38 & 122 & 294 \end{pmatrix}$$

est égale au produit de la matrice triangulaire

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 3 & 0 & 0 \\ 4 & 5 & 6 & 0 \\ 7 & 8 & 9 & 10 \end{pmatrix}$$

avec à sa droite sa transposée L^T :

$$L^T = \begin{pmatrix} 1 & 2 & 4 & 7 \\ 0 & 3 & 5 & 8 \\ 0 & 0 & 6 & 9 \\ 0 & 0 & 0 & 10 \end{pmatrix}$$

Calcul intégrale numérique

Les formules simples (rectangles, trapèzes, simson) utilisent l'interpolation des fonctions à intégrer par des polynômes dont la primitive est connue

$$I = \int_a^b f(x)dx$$

- Formule du rectangle et du point milieu
  - Si ξ est le point d'interpolation : $I(f) - (b-a)f(\xi)$
  - Si $\xi = a$ ou $\xi = b$, l'erreur est $E(f) = \frac{(b-a)^2}{2} f'(\eta)$, $\eta \in [a, b]$: *méthode du rectangle*, d'ordre 0
  - Si $\xi = (a+b)/2$, l'erreur est $E(f) = \frac{(b-a)^3}{24} f''(\eta)$, $\eta \in [a, b]$: *méthode du point milieu*, d'ordre 1

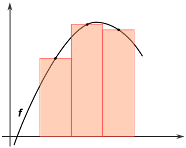


Figure: La surface en rouge représente la valeur de l'intégrale estimée par la méthode du point milieu (wikipedia)

Calcul intégrale numérique

- Formule du trapèze (ordre 1)
  - $I(f) = (b-a)\frac{f(a)+f(b)}{2}$, l'erreur est $E(f) = \frac{(b-a)^3}{12} f''(\eta)$, $\eta \in [a, b]$
- Formule de Simpson (ordre 3)
  - $I(f) = \frac{(b-a)}{6}\left[f(a) + f(\frac{a+b}{2}) + f(b)\right]$, l'erreur est $E(f) = \frac{(b-a)^5}{2880} f^{(4)}(\eta)$, $\eta \in [a, b]$
  - Remarque : comme la méthode du point milieu qui caractérise un polynôme de degré 0 et qui reste exacte pour tout polynôme de degré 1, la méthode de Simpson caractérise un polynôme de degré 2 et reste exacte pour tout polynôme de degré 3.
  - Remarque : si f n'est connue qu'en a et b, le point milieu est calculé par interpolation cubique

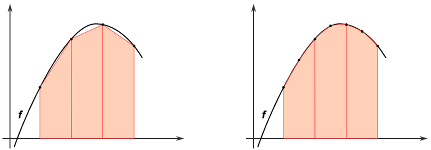


Figure: La surface en rouge représente la valeur de l'intégrale estimée par la méthode des trapèzes (gauche)et de Simpson (droite) (source wikipedia)