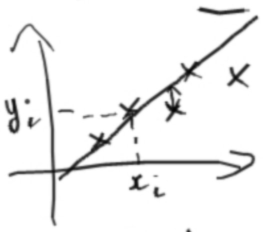
Une application de Gram-Schmidt à la régression linéaire



minimiser la somme $(\text{ordonnée point droite} - \text{ordonnée point})^2$

$$\sum_{i=1}^{n} (ax_i+b-y_i)^2$$

$$= \|\vec{u}-\vec{v}\|^2$$

Il suffit de prendre le produit scalaire usuel sur $\mathbb{R}^n$ (n = nombre de points)

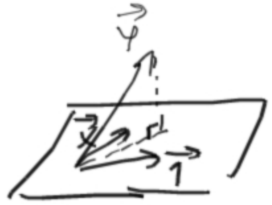
$$\vec{v}\begin{pmatrix} y_1 \\ | \\ y_n \end{pmatrix} = \vec{Y}$$

$$\vec{u}\begin{pmatrix} ax_1+b \\ \vdots \\ ax_n+b \end{pmatrix} = a\vec{X} + b\,\vec{1}$$

$$\vec{X}\begin{pmatrix} x_1 \\ | \\ x_n \end{pmatrix} \quad \vec{1}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

on cherche a et b qui minimisent $\|a\vec{X}+b\vec{1}-\vec{Y}\|^2$

$a\vec{X}+b\vec{1}$ parcourt un plan vectoriel de $\mathbb{R}^n$



La solution est obtenue en calculant le projeté orthogonal du vecteur $\vec{Y}$ sur le plan vectoriel P $\mathrm{Vect}(\vec{1}, \vec{X})$

Gram-Schmidt sur P

$\vec{e}_1 = \vec{1} \quad \vec{e}_2 = \vec{X}$

Normalisation $\vec{e}_1$

$$\|\vec{1}\|^2 = \underbrace{1^2 + \dots + 1^2}_{n \text{ fois}} = n$$

$$\vec{u}_1 = \frac{\vec{1}}{\sqrt{n}}$$

$$pr_{\vec{u}_1}(\vec{e}_2) = \langle \vec{u}_1 | \vec{e}_2 \rangle \vec{u}_1$$

$$= \frac{1}{n} \langle \vec{1} \,|\, \vec{X} \rangle \; \vec{1}$$

$$\langle \vec{1} | \vec{X} \rangle = \left\langle \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \middle| \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \right\rangle$$

$$= x_1 + \dots + x_n$$

$$\frac{1}{n}\langle \vec{1} | \vec{X} \rangle = \frac{x_1 + \dots + x_n}{n}$$

$= \overline{x}$ moyenne des x_i

$$pr_{\vec{u}_1}(\vec{e}_2) = \bar{x}\,\vec{1}$$

$$\vec{f}_2 = \vec{e}_2 - pr_{\vec{u}_1}(\vec{e}_2)$$

$$= \vec{X} - \bar{x}\,\vec{1} = \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix}$$

$$\|\vec{f}_2\|^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\vec{u}_2 = \frac{\vec{X} - \bar{x}\,\vec{1}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Projection de $\vec{Y}$ sur $(\vec{u}_1, \vec{u}_2)$

$$\langle u_1 | \vec{Y} \rangle \vec{u}_1 + \langle \vec{u}_2 | \vec{Y} \rangle \vec{u}_2$$

$$= \frac{1}{n} \langle \vec{1} | \vec{Y} \rangle \vec{1} + \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \langle \vec{X} - \bar{x}\vec{1} | \vec{Y} \rangle (\vec{X} - \bar{x}\vec{1})$$

$$\langle \vec{1} | \vec{Y} \rangle = \sum_{i=1}^{n} y_i$$

$$\langle \vec{X} - \bar{x}\,\vec{1} | \vec{Y} \rangle$$

$$= \sum_{i=1}^{n} (x_i - \bar{x}) y_i$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

car $\bar{y} \sum (x_i - \bar{x}) = 0$

avec $\bar{y} = \frac{\sum y_i}{n}$

$$= \bar{y}\,\vec{1} + \frac{\sigma_{xy}}{\sigma_{xx}} (\vec{X} - \bar{x}\,\vec{1})$$

avec

$$\sigma_{xy} = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\sigma_{xx} = \frac{1}{n} \sum (x_i - \bar{x})^2$$

$$= \underbrace{\frac{\sigma_{xy}}{\sigma_{xx}}}_{a} \vec{X} + \underbrace{(\bar{y} - a\bar{x})}_{b} \vec{1}$$

La droite de régression a pour pente $a = \frac{\sigma_{xy}}{\sigma_{xx}}$

et elle passe par ~~$\bar{x}$ et $\bar{y}$~~ $(\bar{x}, \bar{y})$

$$a\bar{x} + b$$

$$= a\bar{x} + (\bar{y} - a\bar{x})$$

$$= \bar{y}$$