Mesures de tendance centrale et de dispersion d'une série d'intervalles

Marie Chavent, Jérome Saracco

IMB, UMR CNRS 5251, Université Bordeaux1

GREThA, UMR CNRS 5113, Université Montesquieu - Bordeaux IV

Introduction

Classiquement, on décrit une série de n observations réelles $\{x_1, x_2, ..., x_n\}$ par:

- une mesure de tendance centrale comme la moyenne, la médiane....
- une mesure de dispersion comme l'écart-type, l'intervalle inter-quartile.....

Question : Comment décrire une série $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ de n intervalles $\tilde{x}_i = [a_i, b_i] \in I = \{[a, b] \mid a, b \in \mathbb{R}\ ,\ a \leq b\}$.

Idée : Reproduire l'approche géométrique

Introduction

Approche géométrique dans le cas n observations réelles

Principe : choisir $c \in \mathbb{R}$ aussi proche que possible de tous les x_i

Problème d'optimisation :

$$\hat{c} = \arg\min_{c \in \mathbb{R}} S_p(c), \tag{1}$$

avec

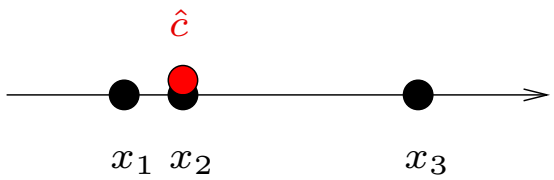
$$S_p(c) = \| \mathbf{x} - \mathbf{c} \|_p = \begin{cases} (\sum_{i=1}^{n} \mid x_i - c \mid^p)^{1/p} & \text{for } p < \infty, \\ \max_{i=1\ldots n} \mid x_i - c \mid & \text{for } p = \infty, \end{cases} \tag{2}$$

$\Rightarrow S_p(\hat{c})$ est la mesure de dispersion associée à la valeur centrale $\hat{c}$.

Introduction

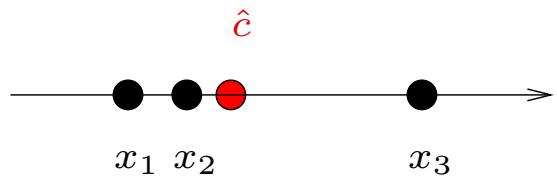
Solutions connues dans le cas n observations réelles :

$p = 1$



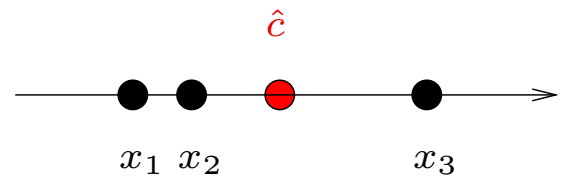
$\hat{c} = x_M$ (médiane)

$p = 2$



$\hat{c} = \bar{x}$ (moyenne)

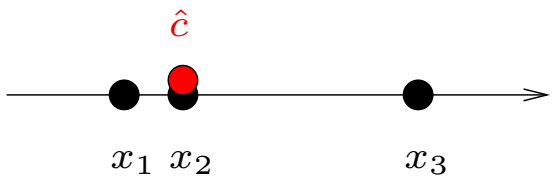
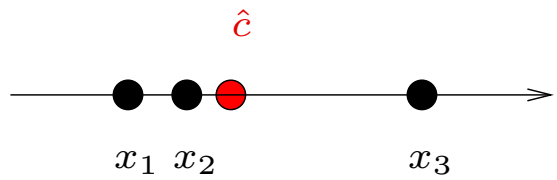
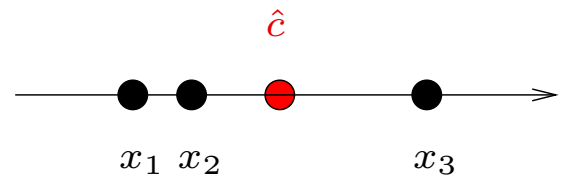
$p = \infty$



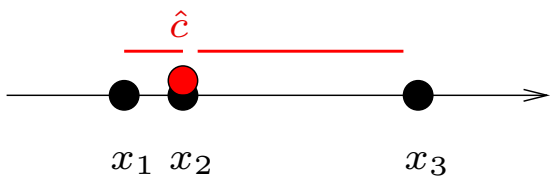
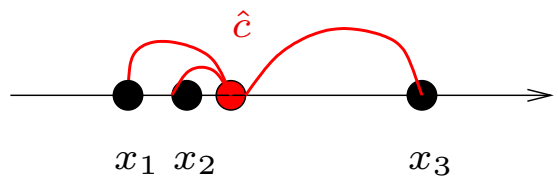
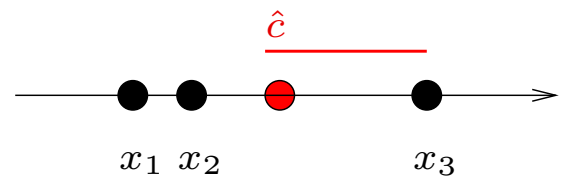
$\hat{c} = x_R$ (midrange)

Introduction

Solutions connues dans le cas n observations réelles :

$p = 1$	$p = 2$	$p = \infty$
		
$\hat{c} = x_M$ (médiane)	$\hat{c} = \bar{x}$ (moyenne)	$\hat{c} = x_R$ (midrange)

Mesures de dispersion associées :

		
$S_1(x_M) = \sum_{i=1}^{n} \mid x_i - x_M \mid$	$S_2(\bar{x}) = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$	$S_\infty(x_R) = \max_i \mid x_i - x_R \mid$
Ecart absolu moyen	Ecart-type	Etendue

Introduction

Approche géométrique dans le cas n intervalles

Principe : choisir $\tilde{c} = [\alpha, \beta]$ aussi proche que possible de tous les $\tilde{x}_i = [a_i, b_i]$

Problème d'optimisation :

$$\hat{\tilde{c}} = \arg\min_{\tilde{c} \in I} \widetilde{S}_p(\tilde{c}), \tag{3}$$

avec en remplacant dans (2) $\mid x_i - c \mid$ par une distance $d(\tilde{x}_i, \tilde{c})$ entre intervalles:

$$\widetilde{S}_p(\tilde{c}) = \begin{cases} (\sum_{i=1}^{n} d(\tilde{x}_i, \tilde{c})^p)^{1/p} & \text{for } p < \infty, \\ \max_{i=1\ldots n} d(\tilde{x}_i, \tilde{c}) & \text{for } p = \infty. \end{cases} \tag{4}$$

$\Rightarrow \widetilde{S}_p(\hat{\tilde{c}})$ est la mesure de dispersion associée à $\hat{\tilde{c}}$.

Quelques distances entre intervalles

Comment définir $d(\tilde{x}_1, \tilde{x}_2)$ avec $\tilde{x}_1 = [a_1, b_1]$ et $\tilde{x}_2 = [a_2, b_2]$?

Première possibilité : Utiliser la distance L_p entre

- les vecteurs $(a_1, b_1)^t$ et $(a_2, b_2)^t$ des bornes inférieures et supérieures des intervalles $\tilde{x}_1$ et $\tilde{x}_2$,
- les vecteurs $(m_1, l_1)^t$ et $(m_2, l_2)^t$ des milieux $m_i = \dfrac{a_i + b_i}{2}$ et des demi-longueurs $l_i = \dfrac{b_i - a_i}{2}$ de $\tilde{x}_1$ et $\tilde{x}_2$.

Quelques distances entre intervalles

Comment définir $d(\tilde{x}_1, \tilde{x}_2)$ avec $\tilde{x}_1 = [a_1, b_1]$ et $\tilde{x}_2 = [a_2, b_2]$?

Seconde possibilité : Utiliser la distance de Hausdorff entre ensemble et qui se simplifie dans le cas de deux intervalles à:

$$d(\tilde{x}_1, \tilde{x}_2) = \max(\mid a_1 - a_2 \mid, \mid b_1 - b_2 \mid) \ , \tag{5}$$

et qui se réecrit facilement

$$d(\tilde{x}_1, \tilde{x}_2) = \mid m_1 - m_2 \mid + \mid l_1 - l_2 \mid \ . \tag{6}$$

$\Rightarrow$ à la fois :

- une distance entre ensembles,
- la distance L_∞ entre les vecteurs $(a_1, b_1)^t$ et $(a_2, b_2)^t$,
- la distance L_1 entre les vecteurs $(m_1, l_1)^t$ et $(m_2, l_2)^t$.

Résultats connus sur les intervalles centraux

Formules explicites pour le calcul de l'intervalle central

(Chavent et Lechevallier 2002, Chavent 2004, De Carvhalo et al. 2006)

Résultat 1 *Dans le cas $p = 1$ d'une combinaison L_1 de distances de Hausdorff, le milieu $\hat{\mu}$ et la demi-longueur $\hat{\lambda}$ de l'intervalle central $\hat{\tilde{c}}$ sont :*

$$\hat{\mu} = median\{m_i \mid i = 1, \ldots, n\}, \quad \hat{\lambda} = median\{l_i \mid i = 1, \ldots, n\}. \quad (7)$$

Résultat 2 *Dans le cas $p = \infty$ d'une combinaison L_∞ de distances de Hausdorff, la borne inférieure $\hat{\alpha}$ et la borne supérieure $\hat{\beta}$ de l'intervalle centrale $\hat{\tilde{c}}$ sont :*

$$\hat{\alpha} = \frac{a_{(n)} - a_{(1)}}{2}, \quad \hat{\beta} = \frac{b_{(n)} - b_{(1)}}{2}, \quad (8)$$

avec $a_{(n)}$ (resp. $b_{(n)}$) la plus grande borne sup (resp. borne inf) et $a_{(1)}$ (resp. $b_{(1)}$) la plus petite borne sup (resp. borne inf).

Résultats connus sur les intervalles centraux

Formules explicites pour le calcul de l'intervalle central

Résultat 3 *Dans le cas $p = 2$ d'une combinaison L_2 de distances L_2 entre les milieux et les demi-longueurs, le milieu $\hat{\mu}$ et la demi-longueur $\hat{\lambda}$ de l'intervalle central $\hat{\tilde{c}}$ sont:*

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} m_i \quad \textit{et} \quad \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} l_i \ .$$

Dans le cas d'une combinaison L_2 de distances L_2 entre les bornes inf et sup, la borne inf et la borne sup de l'intervalle centrale $\hat{\tilde{c}}$ sont :

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} a_i \quad \textit{et} \quad \hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} b_i \ .$$

$\Rightarrow$ Manque le cas $p = 2$ d'une combinaison L_2 de distances de Hausdorff.

Nouveau résultat

Résultat 4 *Dans le cas $p = 2$ d'une combinaison L_2 de distances de Hausdorff, l'intervalle central $\tilde{c}$ qui minimise*

$$\left(\widetilde{S}_2(\tilde{c})\right)^2 = \sum_{i=1}^{n} (\max \mid a_i - \alpha \mid, \mid b_i - \beta \mid)^2 \ . \tag{9}$$

peut être calculé en un nombre fini d'opérations proportionnel à n^3.

Démonstration :

On a

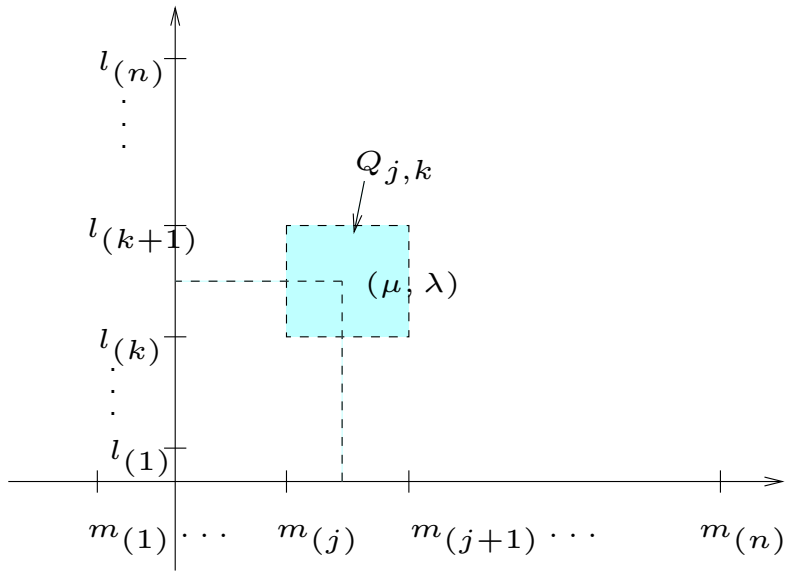
- $\left(\widetilde{S}_2(\tilde{c})\right)^2 = \sum_{i=1}^{n} \max\left((a_i - \alpha)^2, (b_i - \beta)^2\right)$.
- $(a_i - \alpha)^2 - (b_i - \beta)^2 = -4(m_i - \mu)(l_i - \lambda)$.

Donc :

$$\max\left((a_i - \alpha)^2, (b_i - \beta)^2\right) = \begin{cases} (a_i - \alpha)^2 \text{ si } (m_i - \mu)(l_i - \lambda) \leq 0, \\ (b_i - \beta)^2 \text{ si } (m_i - \mu)(l_i - \lambda) \geq 0. \end{cases}$$

Nouveau résultat

On note $(m_{(1)}, \ldots, m_{(n)})$, resp. $(l_{(1)}, \ldots, l_{(n)})$ les milieux et les demi-longueurs ordonnées par ordre croissant



Pour tout (μ, λ) dans un rectangle $Q_{j,k}$, le produit $(m_i - \mu)(l_i - \lambda)$ a un signe donné et ce pour chaque $i = 1, \ldots, n$.

Nouveau résultat

La formule de $\left(\widetilde{S}_2(\tilde{c})\right)^2$ se simplifie donc sur chaque rectangle $Q_{j,k}$:

$$\widetilde{S}_{j,k}(\tilde{c}) = \sum_{i \in I_{a,j,k}} (a_i - \alpha)^2 + \sum_{i \in I_{b,j,k}} (b_i - \beta)^2 \ ,$$

avec :

$$I_{a,j,k} = \left\{ i \in \{1 \dots n\} | \left(m_i - \frac{m_{(j)} + m_{(j+1)}}{2}\right)\left(l_i - \frac{l_{(k)} + l_{(k+1)}}{2}\right) \leq 0 \right\},$$

$$I_{b,j,k} = \left\{ i \in \{1 \dots n\} | \left(m_i - \frac{m_{(j)} + m_{(j+1)}}{2}\right)\left(l_i - \frac{l_{(k)} + l_{(k+1)}}{2}\right) > 0 \right\}.$$

Nouveau résultat

On en déduit que la minimisation de $\left(\widetilde{S}_2(\tilde{c})\right)^2$ sur $\mathbb{R}^2$ est équivalente à la résolution, pour $j,k = 0,1\ldots n$, des $(n+1)^2$ problèmes quadratiques contraints:

$$(\mathrm{P}_{j,k}) \begin{cases} \text{Trouver } (\hat{\alpha}_{j,k}, \hat{\beta}_{j,k}) \text{ qui minimise } \widetilde{S}_{j,k}(\alpha,\beta) \\ \text{sous les contraintes:} \\ 2m_{(j)} \leq \alpha+\beta \leq 2m_{(j+1)} \text{ and } 2l_{(k)} \leq \beta-\alpha \leq 2l_{(k+1)} \end{cases}$$

dont la résolution en un nombre d'opération proportionnel à n est données dans Chavent et Saracco (en révision).

Finalement l'intervalle central $\hat{\tilde{c}} = [\hat{\alpha}, \hat{\beta}]$ est obtenu par:

$$(\hat{\alpha},\hat{\beta}) = \arg \min_{j,k=0,1\ldots,n} \widetilde{S}_{j,k}(\hat{\alpha}_{j,k},\hat{\beta}_{j,k}). \qquad (10)$$

Conclusion

- Applications de ces résultats pour la classification de données intervalles (hypercubes) par la méthode des Nuées dynamiques:
  - définition d'un prototype optimal aussi appelé centrocube.
  - extension simple de ces résultats au cas multidimentionnel en utilisant comme distance entre deux hypercubes $\tilde{\mathbf{x}}_1$ et $\tilde{\mathbf{x}}_2$ une combinaison L_q des distances entre intervalles:

$$D_q(\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2) = \begin{cases} (\sum_{j=1}^{k} d(\tilde{x}_1^j, \tilde{x}_2^j)^q)^{1/q} & \text{for } q < \infty, \\ \max_{j=1\ldots k} d(\tilde{x}_1^j, \tilde{x}_2^j) & \text{for } q = \infty. \end{cases}$$

  - extension difficile et problème ouvert si l'on utilise comme distance la distance de Hausdorff entre deux hypercubes $\tilde{\mathbf{x}}_1$ et $\tilde{\mathbf{x}}_2$ et non pas une combinaison de Hausdorff "uni-dimensionel".

Conclusion

- Applications de ces résultats pour définir une distance normalisée entre vecteurs d'intervalles

$$D_q(\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2) = \begin{cases} (\sum_{j=1}^{k} (\frac{d(\tilde{x}_1^j, \tilde{x}_2^j)}{\widetilde{S}(\hat{\tilde{c}}^j)})^q)^{1/q} & \text{for } q < \infty, \\ \max_{j=1\ldots k} \frac{d(\tilde{x}_1^j, \tilde{x}_2^j)}{\widetilde{S}(\hat{\tilde{c}}^j)} & \text{for } q = \infty. \end{cases}$$

où $\widetilde{S}(\hat{\tilde{c}}^j)$ est la mesure de dispersion associée à l'interval central $\hat{\tilde{c}}^j$. Pour rester cohérent on choisira généralement $q = p$.

Mesures de tendance centrale et de dispersion d'une série d'intervalles

Marie Chavent, Jérome Saracco

IMB, UMR CNRS 5251, Université Bordeaux1

GREThA, UMR CNRS 5113, Université Montesquieu - Bordeaux IV