

Quantiles multivariés par transport optimal et application à la régression quantile

Projet d'expertise en statistiques et probabilités 2020/2021

Théophile Baranger Sébastien Guiroy
Exaucé Luweh Adjim Ngarti

Supervisé par Jérémie Bigot

Université de Bordeaux, Mai 2021

Structure de la présentation

Introduction

Transport optimal et quantiles multivariés

- Transport optimal

- Lien entre transport optimal et quantiles

- Quantiles de Monge-Kantorovich

Application à la régression quantile

- Régression quantile univariée

- Régression quantile vectorielle

- Application aux données ANSUR II

- Application aux données Engel

Structure de la présentation

Introduction

Transport optimal et quantiles multivariés

Transport optimal

Lien entre transport optimal et quantiles

Quantiles de Monge-Kantorovich

Application à la régression quantile

Régression quantile univariée

Régression quantile vectorielle

Application aux données ANSUR II

Application aux données Engel

Introduction

Quantiles en dimension 1

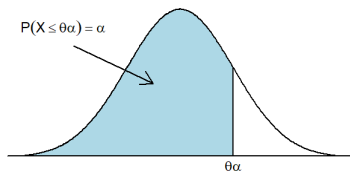
Définition

On considère une variable aléatoire X sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ et sa fonction de répartition F_X continue. La fonction quantile Q est définie pour tout $\alpha \in]0, 1]$ par

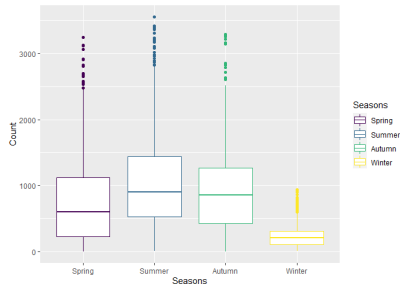
$$Q(\alpha) = F_X^{-1}(\alpha),$$

où F_X^{-1} est l'inverse généralisée de F_X .

Quelques visualisations de quantiles

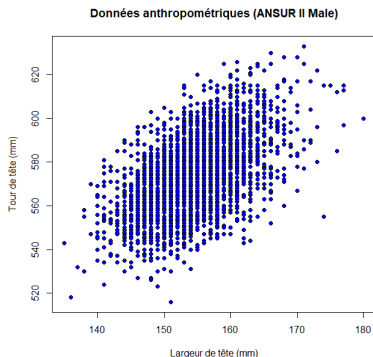


Quantile classique en dimension 1.



Nombre de vélos loués par jour à Séoul en 2019, par saison.

Quantiles en dimensions supérieures



Tour de tête en fonction de la largeur de la tête. (ANSUR II - Male)

Comment généraliser ?

- ▶ Pas de relation d'ordre en dimension $d > 1$,
- ▶ Pas d'extension de la définition d'inverse généralisée,
- ▶ Une extension de la relation d'ordre en dimension supérieure peut être le gradient d'une fonction convexe.

Structure de la présentation

Introduction

Transport optimal et quantiles multivariés

Transport optimal

Lien entre transport optimal et quantiles

Quantiles de Monge-Kantorovich

Application à la régression quantile

Régression quantile univariée

Régression quantile vectorielle

Application aux données ANSUR II

Application aux données Engel

Le transport optimal

Transport de mesures de probabilité

Définition

Soient deux variables aléatoires $X \sim \mu$ et $Y \sim \nu$ définies sur leurs espaces mesurables respectifs \mathcal{X} et \mathcal{Y} . On appelle fonction de transport $T : \mathcal{X} \rightarrow \mathcal{Y}$ une application mesurable telle que

$$T(X) \sim Y.$$

On note alors $T\#\mu = \nu$.

Le transport optimal

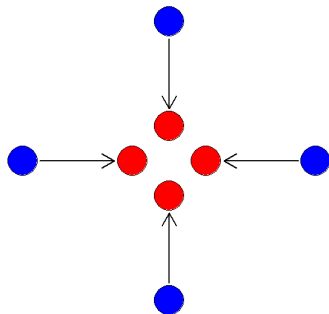
Le problème du transport optimal consiste à résoudre le problème [Peyré and Cuturi, 2020] de minimisation

$$\text{Monge :} \quad \min_T \left\{ \int_{\mathcal{X}} c(x, T(x)) d\mu(x) : T\#\mu = \nu \right\},$$

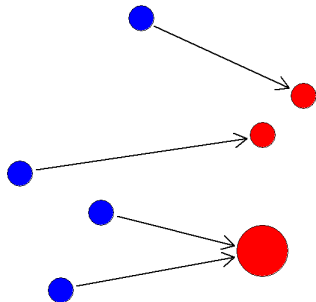
$$\text{Kantorovich :} \quad \min_{\pi \in \mathcal{U}(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y).$$

avec $\mathcal{U}(\mu, \nu)$ l'ensemble des mesures jointes de l'espace produit de lois marginales μ et ν , et $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ une fonction de coût.

Problème de Monge

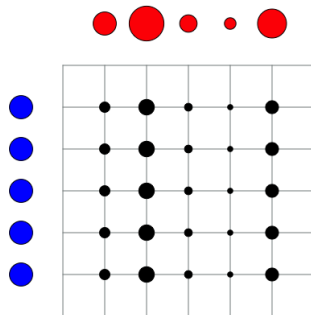


Problème de Monge dans le cas discret. Chaque point de départ est associé à une unique destination.

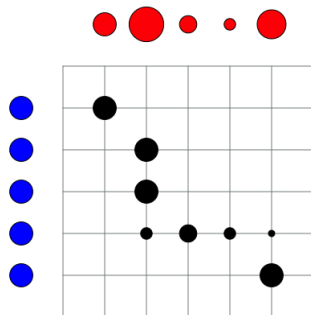


Ici, il y a plus de points de départ que de destinations.

Relaxation de Kantorovich



Solution naïve, en général non-optimale, au problème de Kantorovich.



Solution alternative. Ici, on effectue moins de déplacements, la solution est donc meilleure.

Lien entre transport optimal et quantiles

Théorème d'inversion

Soit X une variable aléatoire de fonction de répartition F_X continue et strictement croissante et $U \sim \mathcal{U}([0, 1])$ Alors la variable aléatoire $F_X^{-1}(U)$ suit la même loi que X ,

$$F_X^{-1}(U) \sim X.$$

Donc par définition du transport de mesure de probabilité,

$$F_X^{-1} \# U = X.$$

Lien entre transport optimal et quantiles

Généralisation en dimension supérieure

Soit X une distribution elliptique à densité radiale et de plein rang,

$$Y := \Sigma^{-1/2}(X - m),$$

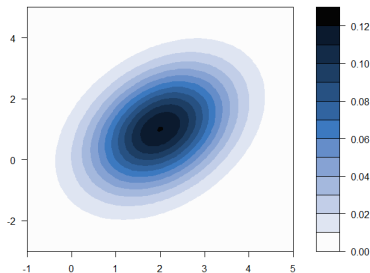
avec m le centre et Σ matrice réelle symétrique définie positive.

Alors Y est une distribution sphérique.

Lien entre transport optimal et quantiles

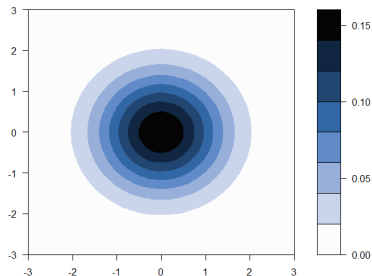
Distributions elliptiques

Distribution elliptique à densité radiale
en dimension 2



Contours de $X \sim \mathcal{N}(\mu, \Sigma)$, avec
 $\mu = (2, 1)$ et $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$.

Distribution sphérique à densité radiale
en dimension 2



Contours de $\mathcal{N}(0, I_2)$, la distribution
résiduelle de X .

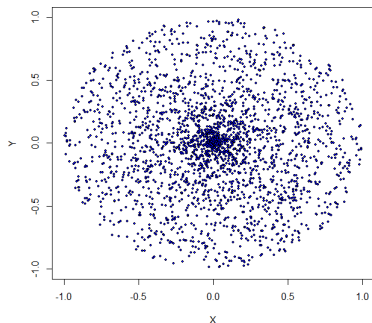
Lien entre transport optimal et quantiles

Uniforme sphérique

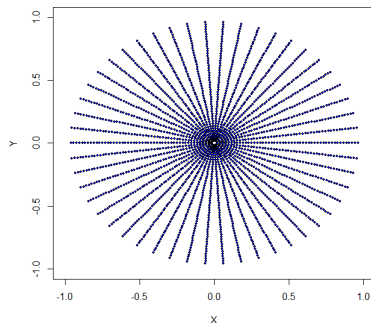
Distribution uniforme sphérique

Notée U_d , c'est la distribution d'un vecteur aléatoire $R \times A$, avec $R \sim \mathcal{U}([0, 1])$ et $A \sim \mathcal{U}(\{x \in \mathbb{R}^d : \|x\| = 1\})$, et $R \perp\!\!\!\perp A$.

Échantillon selon l'uniforme sphérique



Distribution uniforme sphérique empirique



Lien entre transport optimal et quantiles

Fonction de transport

Soit G la fonction de répartition de $\|Y\|$, alors

$$R_P(Y) := \frac{Y}{\|Y\|} G(\|Y\|) \sim U_d.$$

Autrement dit,

$$R_P \# Y = U_d.$$

Les auteurs de [Chernozhukov et al., 2017] définissent la fonction quantile Q_P comme l'inverse de la fonction R_P , et Q_P est le gradient d'une fonction convexe.

Quantiles de Monge-Kantorovich

Cas semi-discret

Théorème (Brenier-McCann)

Soient F et P deux distributions sur \mathbb{R}^d . Si F est absolument continue par rapport à la mesure de Lebesgue sur \mathbb{R}^d , alors pour tout ensemble convexe $\mathcal{U} \subset \mathbb{R}^d$ contenant le support de F , il existe une fonction convexe $\psi : \mathcal{U} \rightarrow \mathbb{R} \cup \{+\infty\}$ telle que $\nabla \psi \# F = P$. Le gradient $\nabla \psi$ de cette fonction existe et est unique F -presque partout.

Quantiles de Monge-Kantorovich

Cas discret-discret

Formellement, on dispose de deux n -uplets $\mathcal{U}_n = \{u_1, \dots, u_n\}$ et $\mathcal{Y}_n = \{y_1, \dots, y_n\}$ de \mathbb{R}^d , et on définit les distributions empiriques

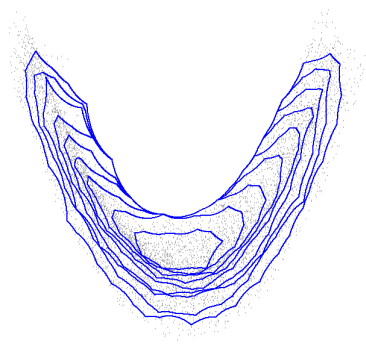
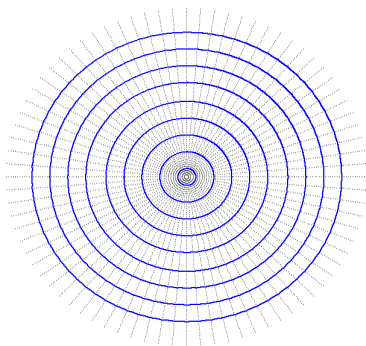
$$F_n = \sum_{j=1}^n \delta_{u_j} / n \quad \text{et} \quad P_n = \sum_{j=1}^n \delta_{y_j} / n,$$

et on cherche à résoudre le problème suivant

$$\sigma^* = \min_{\sigma \in S_n} \sum_{j=1}^n \|u_j - y_{\sigma(j)}\|^2.$$

Quantiles de Monge-Kantorovich

Cas discret-discret

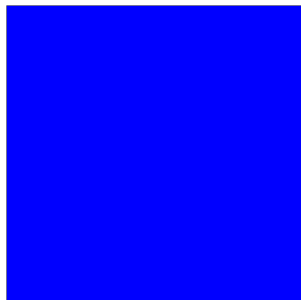


Transport de régions de probabilités (10000 observations), pour les régions de probabilités de niveaux $\tau \in \{0.05, 0.15, \dots, 0.85\}$.

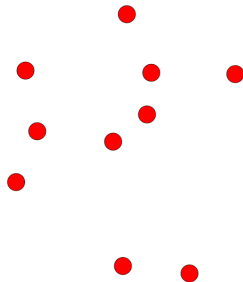
Quantiles de Monge-Kantorovich

Cas semi-discret

On dispose cette fois-ci d'un n -uplet $\mathcal{Y}_n = \{y_1, \dots, y_n\}$ d'observations, et on cherche à transporter une mesure initiale continue vers la mesure empirique des observations.



Mesure de référence continue



Mesure d'intérêt discrète

Quantiles de Monge-Kantorovich

Cas semi-discret

Soient $(y_1, \dots, y_n) \in \mathbb{R}^d$ des observations de notre distribution d'intérêt, $\hat{P}_n = \sum_{j=1}^n \delta_{y_j} / n$ la mesure empirique des données et la distribution de référence F uniforme sphérique. Alors [Chernozhukov et al., 2017] proposent l'estimateur

$$\hat{\psi}_n(u) = \max_{1 \leq k \leq n} \{u^\top y_k - v_k^*\} \quad \forall u \in \mathcal{U}.$$

Où (v_1^*, \dots, v_n^*) minimise $f: \{v_1, \dots, v_n\} \mapsto \int \hat{\psi}_n d\hat{F}_n + \sum_{k=1}^n \frac{v_k}{n}$.

Lemme

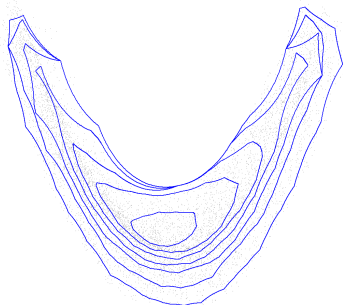
Pour tout $u \in \mathcal{U}$, $\nabla \hat{\psi}_n(u) = y_{k(u)}$, où $k(u) = \operatorname{argmax}_{1 \leq k \leq n} \{u^\top y_k - v_k^*\}$.

Quantiles de Monge-Kantorovich

Cas semi-discret



Découpage de l'espace de départ $[0, 1]^2$, dans le cadre d'un transport vers une mesure discrète de cardinalité 10



Contours estimés sur l'espace d'arrivée des zones de probabilités délimitées sur l'espace de départ (ici, la boule unité en dimension 2)

Structure de la présentation

Introduction

Transport optimal et quantiles multivariés

Transport optimal

Lien entre transport optimal et quantiles

Quantiles de Monge-Kantorovich

Application à la régression quantile

Régression quantile univariée

Régression quantile vectorielle

Application aux données ANSUR II

Application aux données Engel

Régression quantile

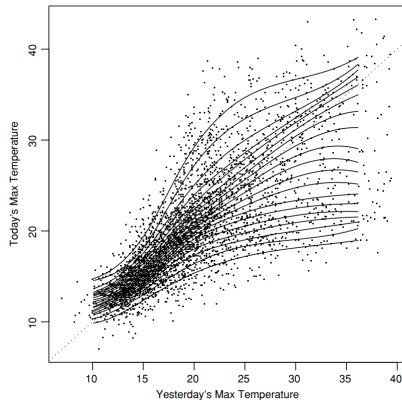
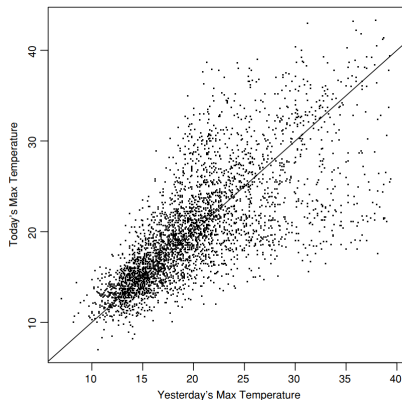
Introduction

Pourquoi la régression quantile ?

- ▶ La régression classique s'intéresse à l'espérance conditionnelle
- ▶ On cherche à obtenir une image plus complète d'un phénomène

Régression quantile

Introduction (suite)



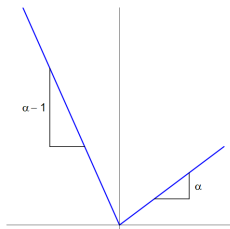
Températures à Melbourne ([Koenker et al., 2005], pp. 51-52)

Régression quantile

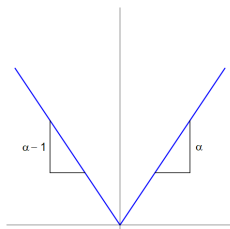
Le quantile comme solution d'un problème d'optimisation

On considère en suivant [Koenker et al., 2005] une fonction de perte définie pour α fixé, par

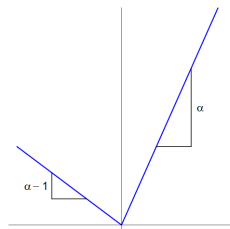
$$\forall x \in \mathbb{R}, \quad \rho_{\alpha}(x) = x(\alpha - \mathbb{1}_{\{x < 0\}}).$$



$\alpha = 0.25$



$\alpha = 0.5$



$\alpha = 0.75$

Régression quantile

Le quantile comme solution d'un problème d'optimisation (suite)

On cherche ensuite à minimiser l'espérance, vue comme une fonction de x^* ,

$$\begin{aligned}\mathbb{E}[\rho_\alpha(X - x^*)] &= \int_{-\infty}^{+\infty} \rho_\alpha(x - x^*) dF(x) \\ &= (\alpha - 1) \int_{-\infty}^{x^*} x - x^* dF(x) + \alpha \int_{x^*}^{+\infty} x - x^* dF(x).\end{aligned}$$

Dont la dérivée (par dérivation sous l'intégrale) s'exprime

$$\begin{aligned}\frac{\partial}{\partial x^*}(\mathbb{E}[\rho_\alpha(X - x^*)]) &= (1 - \alpha) \int_{-\infty}^{x^*} dF(x) - \alpha \int_{x^*}^{+\infty} dF(x) \\ &= \int_{-\infty}^{x^*} dF(x) - \alpha \int_{-\infty}^{+\infty} dF(x) \\ &= F(x^*) - \alpha.\end{aligned}$$

Régression quantile

Le quantile comme solution d'un problème d'optimisation (suite et fin)

La dérivée s'annule lorsque $F(x^*) = \alpha$, et on retrouve ainsi la définition classique de quantile.

Quantile empirique

L'équivalent empirique est donné par

$$x^* = \min_{x \in \mathbb{R}} \sum_{i=1}^n \rho_{\alpha}(y_i - x).$$

Régression quantile

Fondements de la régression quantile

Régression linéaire

On sait que la moyenne empirique est solution du problème

$$\min_{\mu \in \mathbb{R}} \sum_{i=1}^n (y_i - \mu)^2.$$

Ainsi, si l'on pose le modèle linéaire standard

$$\mathbb{E}[Y|X] = \beta^\top X,$$

on estime β en résolvant le problème de minimisation

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - x_i^\top \beta)^2.$$

Régression quantile

Fondements de la régression quantile (suite)

Régression quantile [Koenker and Bassett, 1978]

En posant que le quantile conditionnel d'ordre α satisfait une relation linéaire $\mathcal{Q}_Y(\alpha|X) = X^\top \beta(\alpha)$, l'estimateur $\hat{\beta}(\alpha)$ résout

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \rho_\alpha(y_i - x_i^\top \beta).$$

Ce problème peut se réécrire comme un problème de programmation linéaire [Koenker et al., 2005]

$$\min_{(\beta, u, v) \in \mathbb{R}^p \times \mathbb{R}_+^{2n}} \{ \alpha \mathbb{1}_n^\top u + (1 - \alpha) \mathbb{1}_n^\top v \mid X\beta + u - v = Y \}.$$

Régression quantile

Exemple iid

Modèle de régression

On pose le modèle linéaire suivant

$$y_i = \beta_0 + x_i\beta + \varepsilon_i.$$

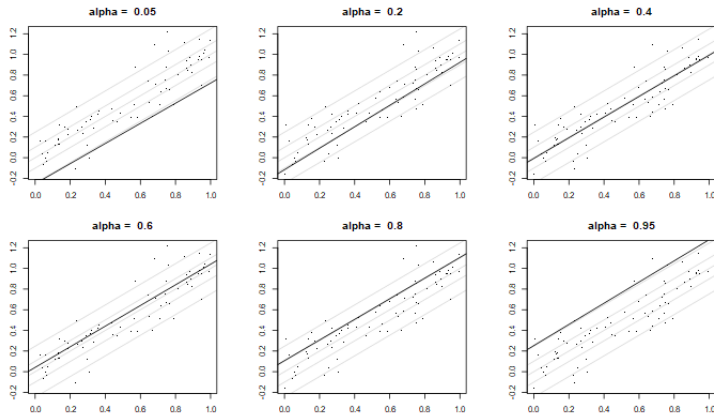
Quantiles conditionnels

Il est clair que les quantiles conditionnels de y sont données par

$$\mathcal{Q}_y(\alpha|x) = \beta_0 + x\beta_1 + F_\varepsilon^{-1}(\alpha).$$

Régression quantile

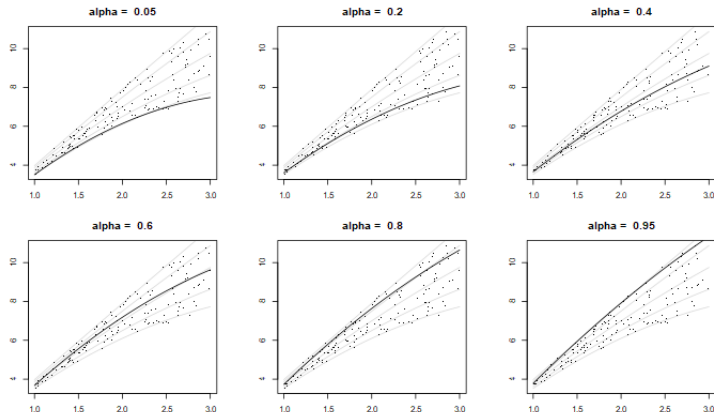
Exemple iid (suite)



Lignes de régression quantile dans le cas d'erreurs identiquement distribuées. Les lignes sont translatées verticalement.

Régression quantile

Exemple non iid



Lignes de régression quantile dans le cas d'erreurs non identiquement distribuées. On remarque que les lignes de régression suivent bien l'allure du nuage de point.

Régression quantile vectorielle

Lien entre régression quantile et transport optimal

Soient X et Y deux variables alors le quantile d'ordre α de Y sachant $X = x$ noté $q_\alpha(x)$ pour $\alpha \in [0, 1]$ est le minimiseur de la fonction

$$\mathbb{E}[\rho_\alpha(Y - q_\alpha(x)) | X = x].$$

Régression quantile vectorielle

Lien entre régression quantile et transport optimal

Soient X et Y deux variables alors le quantile d'ordre α de Y sachant $X = x$ noté $q_\alpha(x)$ pour $\alpha \in [0, 1]$ est le minimiseur de la fonction

$$\mathbb{E}[\rho_\alpha(Y - q_\alpha(x)) | X = x].$$

\downarrow

$$\min_{\beta} \mathbb{E}[(Y - \beta^\top X)^+ + (1 - \alpha)(\beta^\top X - Y)].$$

Et on fait l'hypothèse de linéarité

$$q_\alpha(X) = \beta^\top X.$$

Régression quantile vectorielle

Lien entre régression quantile et transport optimal

Soient X et Y deux variables alors le quantile d'ordre α de Y sachant $X = x$ noté $q_\alpha(x)$ pour $\alpha \in [0, 1]$ est le minimiseur de la fonction,

$$\mathbb{E}[\rho_\alpha(Y - q_\alpha(x)) | X = x].$$

↓

$$\min_{\beta} \mathbb{E}[(Y - \beta^\top X)^+ + (1 - \alpha)(\beta^\top X - Y)].$$

↓

$$\min_{\beta_\alpha} \int_0^1 \mathbb{E}[(Y - \beta_\alpha^\top X)^+ + (1 - \alpha)\beta_\alpha^\top X] d\alpha.$$

Régression quantile vectorielle

Lien entre régression quantile et transport optimal

Soient X et Y deux variables alors le quantile d'ordre α de Y sachant $X = x$ noté $q_\alpha(x)$ pour $\alpha \in [0, 1]$ est le minimiseur de la fonction,

$$\mathbb{E}[\rho_\alpha(Y - q_\alpha(x)) | X = x].$$

↓

$$\min_{\beta} \mathbb{E}[(Y - \beta^\top X)^+ + (1 - \alpha)(\beta^\top X - Y)].$$

↓

$$\min_{\beta_\alpha} \int_0^1 \mathbb{E}[(Y - \beta_\alpha^\top X)^+ + (1 - \alpha)\beta_\alpha^\top X] d\alpha.$$

↓

$$\max_{V_\alpha \geq 0} \left[\int_0^1 \mathbb{E}[V_\alpha Y] d\alpha + \min_{U_\alpha \geq 0, \beta_\alpha} \int_0^1 \mathbb{E}[(1 - V_\alpha)u_\alpha + (1 - \alpha - V_\alpha)\beta_\alpha^\top X] d\alpha \right].$$

Régression quantile vectorielle

Lien entre régression quantile et transport optimal

Soient X et Y deux variables alors le quantile d'ordre α de Y sachant $X = x$ noté $q_\alpha(x)$ pour $\alpha \in [0, 1]$ est le minimiseur de la fonction,

$$\max_{V_\alpha \geq 0} \left[\int_0^1 \mathbb{E}[V_\alpha Y] d\alpha + \min_{U_\alpha \geq 0, \beta_\alpha} \int_0^1 \mathbb{E}[(1 - V_\alpha) U_\alpha + (1 - \alpha - V_\alpha) \beta_\alpha^\top X] d\alpha \right].$$

↓

[Carlier et al., 2020]

$$\begin{aligned} & \max_{(U, X, Y) \sim \pi} \mathbb{E}_\pi[UY] \\ & \text{s.c. } U \sim \mathcal{U}([0, 1]) \\ & (X, Y) \sim \nu \\ & \mathbb{E}[X|Y] = \mathbb{E}[X]. \end{aligned}$$

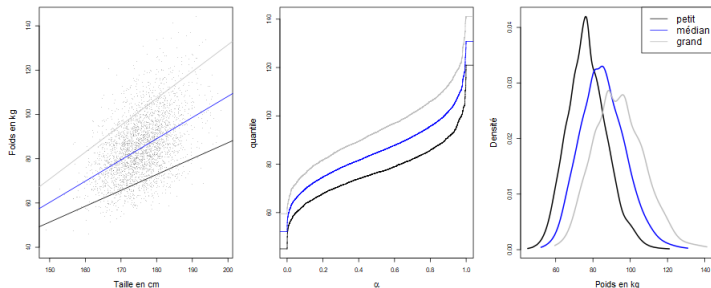
Application aux données ANSUR II

Présentation des données

- ▶ *Anthropometric Survey of US Army Personnel II*
- ▶ 2 jeux de données, 93 mesures relevées, plus de 6.000 sujets
- ▶ Mesures anthropométriques effectuées sur des soldats américains entre 2010 et 2012
- ▶ Une bonne source de données naturellement corrélées

Application aux données ANSUR II

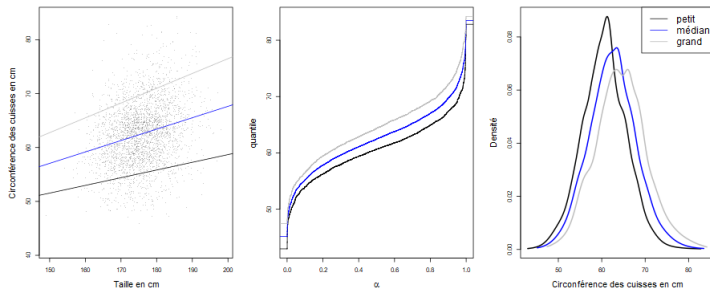
Régression quantile



Estimation de la fonction de quantiles conditionnels du poids (kg) pour trois déciles de taille (cm) (0.1, 0.5, 0.9), que l'on note respectivement « petit », « médian », « grand » (gauche), et densités estimées correspondantes (droite).

Application aux données ANSUR II

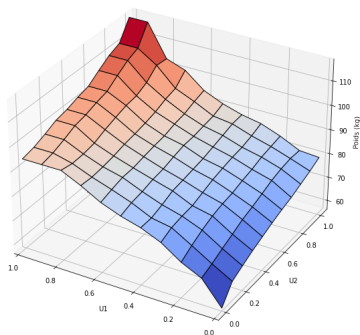
Régression quantile



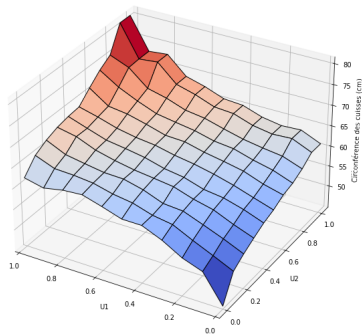
Estimation de la fonction de quantiles conditionnels de la circonférence des cuisses (cm) pour trois déciles de taille (cm) (0.1, 0.5, 0.9), que l'on note respectivement « petit », « médian », « grand » (gauche), et densités estimées correspondantes (droite).

Application aux données ANSUR II

Régression quantile vectorielle



Quantiles conditionnels (taille médiane) pour le poids (kg).



Quantiles conditionnels (taille médiane) pour la circonférence des cuisses (cm).

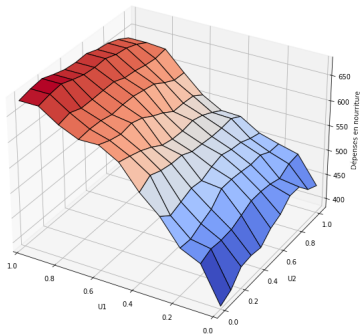
Application aux données Engel

Présentation des données

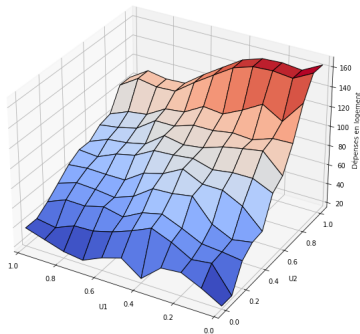
- ▶ Données économiques récoltées par Ernst Engel
- ▶ Dépenses de ménages de classe ouvrière en Belgique en 1857
- ▶ Permet de comparer les dépenses en nourriture, logements, éducation, santé, etc.
- ▶ On s'intéresse à la relation entre dépense en nourriture et dépense pour le logement

Application aux données Engel

Régression quantile vectorielle



Quantiles conditionnels (salaires médians) pour les dépenses en nourriture (francs belges).





Quantiles conditionnels (salaires médians) pour les dépenses en logement (francs belges).

Conclusion

Perspectives futures

- ▶ Version régularisée
- ▶ Construction d'estimateurs
- ▶ Application concrète ?

Références

-  Carlier, G., Chernozhukov, V., De Bie, G., and Galichon, A. (2020).
Vector quantile regression and optimal transport, from theory to numerics.
Empirical Economics.
-  Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017).
Monge–Kantorovich depth, quantiles, ranks and signs.
The Annals of Statistics, 45(1):223 – 256.
-  Koenker, R. and Bassett, G. (1978).
Regression quantiles.
Econometrica, 46(1):33–50.
-  Koenker, R., Cheshier, A., and Jackson, M. (2005).
Quantile Regression.
Econometric Society Monographs. Cambridge University Press.
-  Peyré, G. and Cuturi, M. (2020).
Computational optimal transport.