Multiple-output composite quantile regression via optimal transport

Tengyao Wang

London School of Economics

Statistics Research Showcase

Jun 2024

Collaborator

Xuzhi Yang

Multiple-output linear model

- **Data**: $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} P^{(X,Y)}$ are $\mathbb{R}^p \times \mathbb{R}^d$ random vectors generated from the linear model

$$Y_i = b^* X_i + \epsilon_i,$$

with $b^* \in \mathbb{R}^{d \times p}$ is the regression coefficient of interest, $\mathbb{E}(X_i) = 0$ and the random noise ϵ_i is independent of X_i.

- **Goal**: estimate b^* given data.

Ordinary least squares estimator

- **OLS**: Minimising the residual sum of squares:

$$\hat{b}^{\mathrm{OLS}} := \operatorname*{argmin}_{b \in \mathbb{R}^{d \times p}} \sum_{i=1}^{n} \|Y_i - bX_i\|_2^2.$$

Gauss–Markov: $\hat{b}^{\mathrm{OLS}}$ has minimal variance among all *linear unbiased* estimators.

Ordinary least squares estimator

- **OLS**: Minimising the residual sum of squares:

$$\hat{b}^{\mathrm{OLS}} := \operatorname*{argmin}_{b \in \mathbb{R}^{d \times p}} \sum_{i=1}^{n} \|Y_i - bX_i\|_2^2.$$

  Gauss–Markov: $\hat{b}^{\mathrm{OLS}}$ has minimal variance among all *linear unbiased* estimators.

- But ... one can do a lot better with heavy-tailed noise when we drop the 'linear unbiased' constraint.
- For instance, when $d = 1$ and $\epsilon_i \overset{\mathrm{iid}}{\sim} \mathrm{Cauchy}$, $\hat{b}^{\mathrm{OLS}}$ has infinite variance, but the least absolute deviation regression estimator

$$\hat{b}^{\mathrm{LAD}} := \operatorname*{argmin}_{b \in \mathbb{R}^{1 \times p}} \sum_{i=1}^{n} |Y_i - bX_i|$$

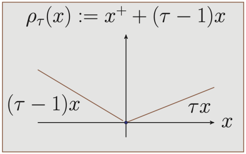
  is still consistent.

Quantile regression for $d = 1$

- LAD regression is a special case of quantile regression (Koenker, 2005).
- When $d = 1$, for any fixed quantile level $\tau \in (0, 1)$, the **quantile regression estimator** is defined as

$$(\hat{b}^{\mathrm{QR}_\tau}, \hat{q}_\tau) := \operatorname*{argmin}_{b \in \mathbb{R}^{1\times p}, q_\tau \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(Y_i - bX_i - q_\tau),$$

where $\rho_\tau(x) = \tau x^+ + (1-\tau)x^- = x^+ + (\tau - 1)x$ is the 'check loss'.



- Under regularity conditions,

$$\sqrt{n}(\hat{b}^{\mathrm{QR}_\tau} - b^*) \xrightarrow{\mathrm{d}} N\biggl(0, \frac{\tau(1-\tau)}{f_\epsilon^2(q_\tau^*)} \Sigma_{\mathrm{x}}^{-1}\biggr),$$

where $\Sigma_{\mathrm{x}} = \mathrm{cov}(X_1)$ and $f_\epsilon(q_\tau^*)$ is the density of ϵ_1 at its τ-quantile.

Composite quantile regression for $d = 1$

- $\hat{b}^{\mathrm{QR}_\tau}$ is $\sqrt{n}$-consistent when ϵ_1 has nonvanishing density at its τ-quantile, though its efficiency can be arbitrarily small.
- The idea of **composite quantile regression** (Zou and Yuan, 2008) is to use multiple quantiles: setting $\tau_k = k/(K+1)$ or $k = 1, \ldots, K$, define

$$\hat{b}^{\mathrm{CQR}} = \operatorname*{argmin}_{b \in \mathbb{R}^{1 \times p}} \min_{q_1 < \cdots < q_K} \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(Y_i - bX_i - q_k).$$

- $\hat{b}^{\mathrm{CQR}}$ has asymptotic variance at most $e\pi/6 \approx 1.4$ times that of OLS estimator and can be much more efficient when noise is heavy-tailed.

What to do when $d \geq 2$?

- Coordinatewise (composite) quantile regression?
- Multivariate generalisation of the quantiles and check functions
  - Projected/directional quantiles (Paindevaine and Šiman, 2011)
  - Spatial quantiles (Chaudhuri, 1996)

What to do when $d \geq 2$?

- Coordinatewise (composite) quantile regression?
- Multivariate generalisation of the quantiles and check functions
  - Projected/directional quantiles (Paindevaine and Šiman, 2011)
  - Spatial quantiles (Chaudhuri, 1996)

- We take a different perspective — recasting the composite quantile regression into an **optimal transport** problem

A very brief introduction to OT

- Given p.m. P and Q on $\mathcal{X}$, the squared **2-Wasserstein distance** $\mathcal{W}_2^2(P,Q)$ is the minimum cost of moving mass from P into Q.
- When P and Q are both empirical measures of n points, this specialises to the **assignment problem**.
- Formally, any transport is a joint distribution (coupling) π on $\mathcal{X} \times \mathcal{X}$ with marginals P and Q and the optimal transport solves the optimisation

$$\pi^* = \operatorname*{argmin}_{\pi \in \mathcal{C}(P,Q)} \mathbb{E}_{(X,Y)\sim\pi} \|X - Y\|^2$$

A very brief introduction to OT

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

- Given p.m. P and Q on $\mathcal{X}$, the squared **2-Wasserstein distance** $\mathcal{W}_2^2(P, Q)$ is the minimum cost of moving mass from P into Q.
- When P and Q are both empirical measures of n points, this specialises to the **assignment problem**.
- Formally, any transport is a joint distribution (coupling) π on $\mathcal{X} \times \mathcal{X}$ with marginals P and Q and the optimal transport solves the optimisation

$$\pi^* = \operatorname*{argmin}_{\pi \in \mathcal{C}(P,Q)} \mathbb{E}_{(X,Y)\sim\pi} \|X - Y\|^2$$

- The **Monge–Kantorovich duality**:

$$\min_{\pi \in \mathcal{C}(P,Q)} \mathbb{E}_\pi \|X - Y\|^2 = \max_{\phi(x)+\psi(y) \le \|x-y\|^2} \mathbb{E}_P \phi(X) + \mathbb{E}_Q \psi(Y).$$

The dual solutions ϕ, ψ satisfies that $x \mapsto x^2/2 - \phi(x)$ and $y \mapsto y^2/2 - \psi(y)$ are convex functions.

CQR in population formula

- Assume infinite data and let $K \to \infty$, then the CQR objective becomes

$$\min_{b \in \mathbb{R}^{1\times p}} \min_{q \in \mathcal{M}} \mathbb{E}\left\{ \int_0^1 \rho_\tau(Y - bX - q(\tau))\, d\tau \right\},$$

where $\mathcal{M}$ denotes the set of increasing functions on $\mathbb{R}$.

CQR in population formula

- Assume infinite data and let $K \to \infty$, then the CQR objective becomes

$$\min_{b \in \mathbb{R}^{1\times p}} \min_{q \in \mathcal{M}} \mathbb{E}\left\{ \int_0^1 \rho_\tau(Y - bX - q(\tau))\, d\tau \right\},$$

where $\mathcal{M}$ denotes the set of increasing functions on $\mathbb{R}$.

- Let $U \sim \mathrm{Unif}[0,1]$ and $\phi(t) = \int_0^t q(\tau)\, d\tau$, we can rewrite

$$\begin{aligned}
\min_{q \in \mathcal{M}} \mathbb{E}\left\{ \int_0^1 \rho_\tau(Y - bX - q(\tau))\, d\tau \right\} + \frac{1}{2}\mathbb{E}(Y) \\
&= \min_{q \in \mathcal{M}} \mathbb{E}\left\{ \int_0^1 (Y - bX - q(\tau))^+\, d\tau + \int_0^1 (1-\tau)q(\tau)\, d\tau \right\} \\
&= \min_{q \in \mathcal{M}} \mathbb{E}\left\{ \max_{t \in [0,1]} \int_0^t (Y - bX - q(\tau))\, d\tau + \int_0^U q(\tau)\, d\tau \right\} \\
&= \min_{\phi \text{ convex}} \left\{ \mathbb{E} \max_{t \in [0,1]} \big(t(Y - bX) - \phi(t)\big) + \mathbb{E}\phi(U) \right\}
\end{aligned}$$

CQR in population formula

- Assume infinite data and let $K \to \infty$, then the CQR objective becomes

$$\min_{b \in \mathbb{R}^{1\times p}} \min_{q \in \mathcal{M}} \mathbb{E}\left\{ \int_0^1 \rho_\tau(Y - bX - q(\tau))\, d\tau \right\}.$$

- Let $U \sim \mathrm{Unif}[0,1]$ and $\phi(t) = \int_0^t q(\tau)\, d\tau$, we can rewrite

$$\begin{aligned} &\min_{q \in \mathcal{M}} \mathbb{E}\left\{ \int_0^1 \rho_\tau(Y - bX - q(\tau))\, d\tau \right\} + \frac{1}{2}\mathbb{E}(Y) \\ &= \min_{\phi \text{ convex}} \left\{ \mathbb{E} \max_{t \in [0,1]} \big(t(Y - bX) - \phi(t) \big) + \mathbb{E}\,\phi(U) \right\} \end{aligned}$$

CQR in population formula

- Assume infinite data and let $K \to \infty$, then the CQR objective becomes

$$\min_{b \in \mathbb{R}^{1 \times p}} \min_{q \in \mathcal{M}} \mathbb{E}\left\{ \int_0^1 \rho_\tau(Y - bX - q(\tau))\, d\tau \right\}.$$

- Let $U \sim \mathrm{Unif}[0,1]$ and $\phi(t) = \int_0^t q(\tau)\, d\tau$, we can rewrite

$$\begin{aligned} &\min_{q \in \mathcal{M}} \mathbb{E}\left\{ \int_0^1 \rho_\tau(Y - bX - q(\tau))\, d\tau \right\} + \frac{1}{2}\mathbb{E}(Y) \\ &= \min_{\phi \text{ convex}} \left\{ \mathbb{E} \max_{t \in [0,1]} \big(t(Y - bX) - \phi(t)\big) + \mathbb{E}\,\phi(U) \right\} \\ \text{(L–F duality)} \quad &= \min_{\phi \text{ convex}} \left\{ \mathbb{E}\,\phi^*(Y - bX) + \mathbb{E}\,\phi(U) \right\} \end{aligned}$$

- Assume infinite data and let $K \to \infty$, then the CQR objective becomes

$$\min_{b \in \mathbb{R}^{1\times p}} \min_{q \in \mathcal{M}} \mathbb{E}\biggl\{ \int_0^1 \rho_\tau(Y - bX - q(\tau))\, d\tau \biggr\}.$$

- Let $U \sim \mathrm{Unif}[0,1]$ and $\phi(t) = \int_0^t q(\tau)\, d\tau$, we can rewrite

$$\begin{aligned}
&\min_{q \in \mathcal{M}} \mathbb{E}\biggl\{ \int_0^1 \rho_\tau(Y - bX - q(\tau))\, d\tau \biggr\} + \frac{1}{2}\mathbb{E}(Y) \\
&= \min_{\phi \text{ convex}} \Bigl\{ \mathbb{E} \max_{t\in[0,1]} \bigl(t(Y - bX) - \phi(t)\bigr) + \mathbb{E}\,\phi(U) \Bigr\} \\
\text{(L–F duality)} \quad &= \min_{\phi \text{ convex}} \Bigl\{ \mathbb{E}\,\phi^*(Y - bX) + \mathbb{E}\,\phi(U) \Bigr\} \\
\text{(M–K duality)} \quad &= \frac{1}{2}\bigl\{ -\mathcal{W}_2^2(Y - bX, U) + \mathbb{E}(Y - bX)^2 + \mathbb{E}\,U^2 \bigr\} \\
&= \sup_{\pi \in \mathcal{C}(P^{Y-bX}, P^U)} \mathbb{E}\,\langle Y - bX, U \rangle =: \langle\!\langle Y - bX, U \rangle\!\rangle_{\mathcal{W}_2}.
\end{aligned}$$

CQR in population formula

- Assume infinite data and let $K \to \infty$, then the CQR objective becomes

$$\min_{b \in \mathbb{R}^{1\times p}} \min_{q \in \mathcal{M}} \mathbb{E}\left\{ \int_0^1 \rho_\tau(Y - bX - q(\tau))\, d\tau \right\}.$$

- Let $U \sim \mathrm{Unif}[0,1]$ and $\phi(t) = \int_0^t q(\tau)\, d\tau$, we can rewrite

$$\begin{aligned}
& \min_{q \in \mathcal{M}} \mathbb{E}\left\{ \int_0^1 \rho_\tau(Y - bX - q(\tau))\, d\tau \right\} + \frac{1}{2}\mathbb{E}(Y) \\
&= \min_{\phi \text{ convex}} \left\{ \mathbb{E} \max_{t \in [0,1]} \big(t(Y - bX) - \phi(t)\big) + \mathbb{E}\,\phi(U) \right\} \\
\text{(L–F duality)} \quad &= \min_{\phi \text{ convex}} \left\{ \mathbb{E}\,\phi^*(Y - bX) + \mathbb{E}\,\phi(U) \right\} \\
\text{(M–K duality)} \quad &= \frac{1}{2}\left\{ -\mathcal{W}_2^2(Y - bX, U) + \mathbb{E}(Y - bX)^2 + \mathbb{E}\,U^2 \right\} \\
&= \sup_{\pi \in \mathcal{C}(P^{Y-bX}, P^U)} \mathbb{E}\,\langle Y - bX, U \rangle =: \langle\!\langle Y - bX, U \rangle\!\rangle_{\mathcal{W}_2}.
\end{aligned}$$

- The population formulation of CQR

$$b^* = \operatorname*{argmin}_{b \in \mathbb{R}^{1\times p}} \langle\!\langle Y - bX, U \rangle\!\rangle_{\mathcal{W}_2}$$

 has an immediate generalisation to multivariate output.

- For any $P^\epsilon, P^U \in \mathcal{P}_2(\mathbb{R}^d) \cap \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ and P^X is not a point mass, b^* uniquely solves the population MCQR objective:

$$b^* = \operatorname*{argmin}_{b \in \mathbb{R}^{d\times p}} \mathcal{L}(b), \quad \text{where } \mathcal{L}(b) := \langle\!\langle Y - bX, U \rangle\!\rangle_{\mathcal{W}_2}.$$

- **MCQR estimator**: given $(X_1, Y_1), \ldots, (X_n, Y_n)$, draw $U_1, \ldots, U_n \sim N_d(0, I_d)$, we define

$$\hat{b} = \hat{b}^{\mathrm{MCQR}} \in \operatorname*{argmin}_{b \in \mathbb{R}^{d\times p}} \mathcal{L}_n(b), \quad \text{where } \mathcal{L}_n(b) := \langle\!\langle P_n^{Y-bX}, P_n^U \rangle\!\rangle_{\mathcal{W}_2}$$

Relation to Monge–Kantorovich quantiles

- The optimal transport coupling between P^{Y-bX} and P^U induces maps F and Q such that $F(Y - bX) \sim P^U$ and $Q(U) \sim P^{Y-bX}$.
- F and Q are known as the **Monge–Kantorovich rank and quantile** functions of P^{Y-bX}.
- These are multivariate generalisations of the ranks and quantiles proposed by Chernozhukov et al. (2017) and Hallin et al. (2021).
- M–K ranks and quantiles have found applications in distribution-free nonparametric statistical inference (Ghosal and Sen, 2022)

Computing the estimator

- For a fixed b, computing $\mathcal{L}_n(b) = \langle\!\langle P_n^{Y-bX}, P_n^U \rangle\!\rangle_{\mathcal{W}_2}$ amounts to an assignment problem. Let $\mathcal{A}$ be the class of assignment matrices.
- Writing $\boldsymbol{U} = (U_1, \ldots, U_n)^\top$, $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, we have
$$\min_{b \in \mathbb{R}^{d \times p}} \mathcal{L}_n(b) = \min_{b \in \mathbb{R}^{d \times p}} \max_{A \in \mathcal{A}} \mathrm{Tr}\big(\boldsymbol{U}^\top A(\boldsymbol{Y} - \boldsymbol{X} b^\top)\big)$$
- Easier to solve the dual problem:
$$\max_{A \in \mathcal{A}} \mathrm{Tr}(\boldsymbol{U}^\top A) \quad \text{s.t. } \boldsymbol{U}^\top A \boldsymbol{X} = 0,$$
by standard LP solvers.

Theoretical guarantees

- We assume that P^ϵ is has a Lebesgue density and P^X is an elliptical distribution.
- **Polynomial-tailed noise**: suppose that P^X and P^ϵ both have finite ℓ-th moment ($\ell > 2$), then with probability at least $1 - \frac{4}{\log n}$, the MCQR estimator satisfies

$$\|\hat{b}^{\mathrm{MCQR}} - b^*\|_\Sigma^2 \wedge 1 \lesssim_{d,p,\log n} n^{-1/4} + n^{-1/\max(d,p)} + n^{-(\ell-2)/(2\ell)}.$$

Theoretical guarantees

- We assume that P^ϵ is has a Lebesgue density and P^X is an elliptical distribution.
- **Polynomial-tailed noise**: suppose that P^X and P^ϵ both have finite ℓ-th moment ($\ell > 2$), then with probability at least $1 - \frac{4}{\log n}$, the MCQR estimator satisfies

$$\|\hat{b}^{\mathrm{MCQR}} - b^*\|_\Sigma^2 \wedge 1 \lesssim_{d,p,\log n} n^{-1/4} + n^{-1/\max(d,p)} + n^{-(\ell-2)/(2\ell)}.$$

- **Sub-Weibull-tailed noise**: Suppose $\Sigma^{-1/2}X_1$ is (σ_1, α)-sub-Weibull and ϵ_1 is (σ_2, β)-sub-Weibull, i.e.

$$\mathbb{E}\, e^{(\|\Sigma^{-1}X_1\|/\sigma_1)^\alpha/2} \leq 2 \quad \text{and} \quad \mathbb{E}\, e^{(\|\epsilon_1\|/\sigma_2)^\beta/2} \leq 2,$$

and the density of ϵ_1 satisfies $f_\epsilon(u) \geq \gamma_1 e^{-\gamma_2\|u\|_2^2}$ for $\|u\| \geq 1$, then with probability at least $1 - \frac{33}{\log n}$, we have

$$\|\hat{b}^{\mathrm{MCQR}} - b^*\|_\Sigma^2 \wedge 1 \lesssim_{d,\log n} \sqrt{\frac{p}{n}} + n^{-2/d}.$$

Proof sketch

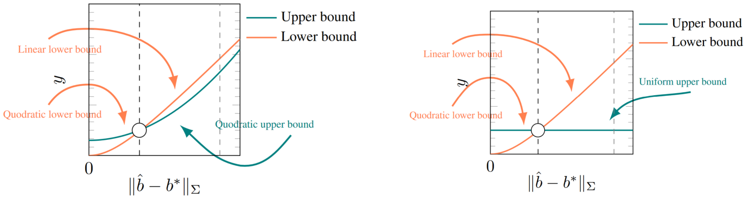
- **Basic inequality**:

$$\mathcal{L}(\hat{b}) - \mathcal{L}(b^*) \leq \mathcal{L}(\hat{b}) - \mathcal{L}_n(\hat{b}) + \mathcal{L}_n(b^*) - \mathcal{L}(b^*).$$

- To control the LHS, we have the following inequality: for random vectors $Z \perp\!\!\!\perp \epsilon$ in $\mathbb{R}^d$ with finite second moment and $U \sim N_d(0, I_d)$, we have

$$\langle\langle Z + \epsilon, U \rangle\rangle^2_{\mathcal{W}_2} \geq \langle\langle Z, U \rangle\rangle^2_{\mathcal{W}_2} + \langle\langle \epsilon, U \rangle\rangle^2_{\mathcal{W}_2}.$$

- To control the RHS, we use bounds for distances between empirical and population 2-Wasserstein distances (Fournier and Guillin, 2015).



Empirical performance of MCQR estimator

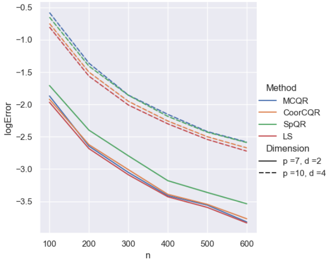
- The noise ϵ_i's generated from one of the following distributions:
  (i) $\epsilon_i \sim \mathcal{N}(0, I_d)$
  (ii) $\epsilon_i \sim t_2(0, I_d)$ follows a multivariate t_2 distribution
  (iii) ϵ_i has each marginal distributed with $\mathrm{Pareto}(-2, 2, 1)$ and the same copula as $\mathcal{N}(0, \Sigma')$, where $\Sigma' = (0.9^{|i-j|})_{i,j} \in \mathbb{R}^{d \times d}$
  (iv) ε follows a centered Banana-shaped distribution.

- We compare the average loss (matrix Mahalanobis norm) of MCQR estimator against
  - Coordinatewise CQR estimator (CoorCQR)
  - Spatial quantile regression estimator (SpQR)
  - OLS estimator

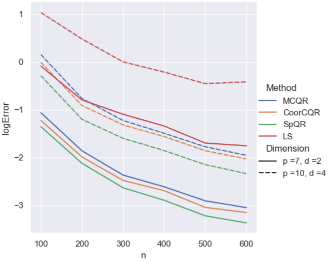
Empirical performance of MCQR estimator

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

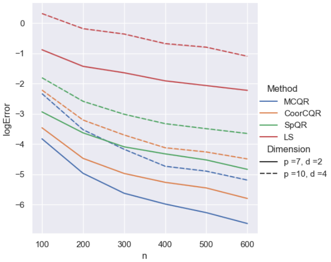
Gaussian noise



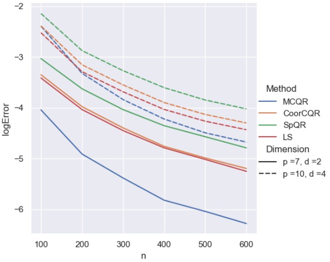
Multivariate t_2



Pareto copula

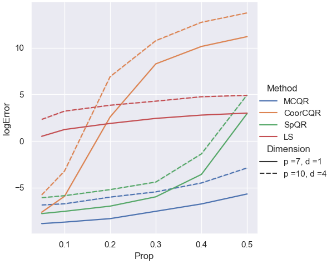


Banana-shaped

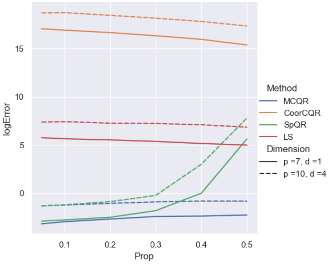


- We also investigate the empirical performance of MCQR in the presence of outlier contamination. Here, we consider two cases of δ-contaminated noise, for some $\delta \in (0,1)$:
  - (i) $\epsilon \sim (1-\delta)P_1 + \delta P_2$; here P_1 is a Pareto copula as before and P_2 is a heavier-tailed location-shifted Pareto copula with marginals distributed as $\mathrm{Pareto}(10, 2, 10)$.
  - (ii) $\varepsilon \sim (1-\epsilon)\mathcal{N}(0, I_d) + \epsilon\mathcal{N}(100, I_d)$

Pareto contamination



Gaussian contamination



Summary

- CQR optimisation has a natural OT interpretation
- This allows a multivariate generalisation
- Current theoretical control is likely suboptimal
- Empirical performance is very promising

Main reference:

- Yang, X. and Wang, T. (2024) Multiple-output composite quantile regression through an optimal transport lens. *COLT 2024.*

Thank you!

References

- Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.*, **91**, 862–872.
- Chernozhukov, V., Galichon, A., Hallin, M. and Henry, M. (2017). Monge-Kantorovich depth, quantiles, ranks and signs. *Ann. Statist.*, **45**, 223–256.
- Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, **162**, 707–738.
- Ghosal, P. and Sen, B. (2022). Multivariate ranks and quantiles using optimal transport: consistency, rates and nonparametric testing. *Ann. Statist.*, **50**, 1012–1037.
- Hallin, M., Del Barrio, E., Cuesta-Albertos, J. and Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: a measure transportation approach. *Ann. Statist.*, **49**, 1139–1165.
- Koenker, R. (2005). *Quantile regression.* Cambridge University Press, Cambridge.
- Paindaveine, D. and Šiman, M. (2011). On directional multiple-output quantile regression. *J. Multivariate Anal.*, **102**, 193–212.
- Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.*, **36**, 1108–1126.