MULTIVARIATE QUANTILES, OPTIMAL TRANSPORT

Gauthier THURIN

Advisor : Jérémie BIGOT
Co-advisor : Bernard BERCU

Univ. Bordeaux, CNRS, INRIA, Bordeaux INP, IMB, UMR 5251, F-33400 Talence, France

Monge-Kantorovich quantiles are defined in an intuitive fashion through the optimal transportation from a reference distribution. My thesis is dedicated to their study, as they benefit from most of sought-after properties. The adaptivity of these concepts makes them attractive for building statistical tools, and has concentrated much attention in the recent years. My PhD project partly belongs to this line of work, as we proposed multivariate definitions of superquantiles, expected shortfalls and related risk measures. In addition, we focused on the estimation of a regularized version of the Monge-Kantorovich quantile function, via a stochastic algorithm with almost-sure convergence guarantees.

Ordering multivariate data

Consider a probability distribution ν supported on $\mathbb{R}^d$. In the scalar case $d = 1$, the quantile function of ν is nothing else than the generalized inverse F_ν^{-1} of the cumulative distribution function F_ν. This heavily relies on the left-to-right ranking of samples from ν. However, in the multi-dimensional case $d > 1$, the lack of a canonical ordering precludes a consensual notion of quantiles, see [1, 2] and references therein. The key ingredient for a concept of quantiles is thus the ordering it provides for observations sampled from ν.

Departing from what is known, F_ν^{-1} appears to be the unique solution of the optimal transport (OT) problem between the uniform $\mu = U([0,1])$ and ν. By stating this as a definition when $d > 1$, the authors of [1] generalized univariate quantiles to define the Monge-Kantorovich (MK) quantile function of ν as

$$\mathbf{Q} = \underset{T:T_\#\mu=\nu}{\operatorname{argmin}} \mathbb{E}_{X\sim\mu}\left(\|X - T(X)\|^2\right),$$

where the push-forward constraint $T_\#\mu = \nu$ means

$$\int_{\mathcal{X}} g\, d\nu = \int_{T^{-1}(\mathcal{X})} g \circ T\, d\mu.$$

A suitable choice of reference measure μ is the one of the random vector $R\Phi$, where R and Φ are independent and drawn uniformly from $[0,1]$ and the unit hypersphere $\mathbb{S}^{d-1} = \{\varphi \in \mathbb{R}^d : \|\varphi\| = 1\}$, respectively. With this choice, the unit balls $\mathbb{B}(0,\alpha)$ have μ-probability α, and the regions $\mathbf{Q}(\mathbb{B}(0,\alpha))$ are nested and have ν-probability α, and thus are appropriate candidates for quantile regions, [1, 2]. Descriptive plots in the spirit of multivariate boxplots are given in Figure 1.1 for a banana-shaped distribution ν. The fact that the random vector $\mathbf{Q}(U)$ follows ν as soon as U is sampled from μ is crucial. It allows to claim that $\mathbf{Q}$ is able to capture all the available information. Roughly, in Figure 1.1, the same amount of mass is contained in the regions delimited by two red curves, and the blue curves encapsulate quantile regions indexed by their probability contents.

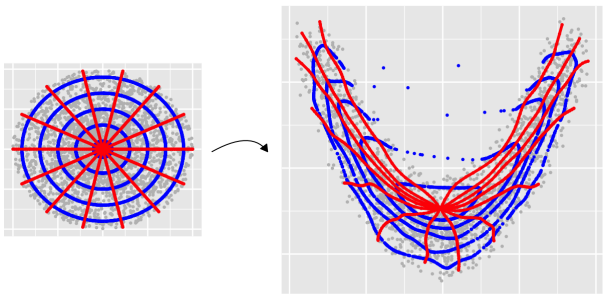

FIGURE 1.1 – *For any distribution, on the right, a transported ordering from some reference measure, on the left.*

Statistical tools

In $\mathbb{R}$, the superquantiles and expected shortfalls complement the information given by the quantiles. By focusing on which univariate features we aim to extend, we defined in [3] the MK superquantile function

$$S(u) = \frac{1}{1 - \|u\|} \int_{\|u\|}^1 \mathbf{Q}(t\frac{u}{\|u\|})dt,$$

and the MK expected shortfall function

$$E(u) = \frac{1}{\|u\|} \int_0^{\|u\|} \mathbf{Q}(t\frac{u}{\|u\|})dt.$$

The counterpart of Figure 1.1 by using S and E instead of $\mathbf{Q}$ induces the descriptive plots of Figure 1.2. These functions describe central and peripheral areas of point clouds, and they can be shown to characterize random vectors and their convergence in distribution, [3].

Theorem. *Let ν_1 and ν_2 be probability distributions on $\mathbb{R}^d$, with respective MK superquantile and expected shortfall functions given by S_1, E_1 and S_2, E_2. Then,*

$$\nu_1 = \nu_2 \Leftrightarrow S_1 = S_2 \ \mu\text{-a.e.} \Leftrightarrow E_1 = E_2 \ \mu\text{-a.e.}$$

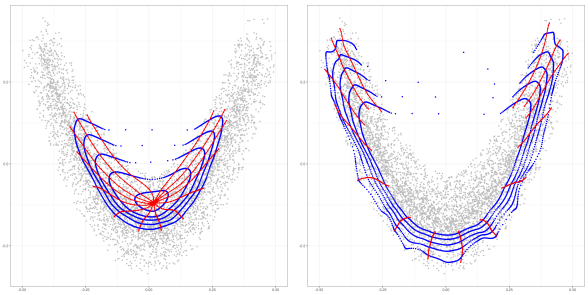

FIGURE 1.2 – *Descriptive plots, E on the left, S on the right.*

Theorem. *Let ν_n and ν be such that their respective MK quantile functions $\mathbf{Q}_n$ and $\mathbf{Q}$ are continuous on $\mathbb{B}(0,1)\backslash 0$. Then the weak convergence of ν_n towards ν is equivalent to*

$$\forall u \in \mathbb{B}(0,1)\backslash 0, \lim_{n\to+\infty} E_n(u) = E(u)$$

and, for any compact $K \subset \mathbb{B}(0,1)\backslash 0$,

$$\lim_{n\to+\infty} \sup_{u\in K} \|E_n(u) - E(u)\| = 0.$$

Besides, if one assumes that, for every $n \in \mathbb{N}$ and X_n drawn from ν_n, there exists a random vector Z with $\mathbb{E}[Z\ln(Z)] < +\infty$ such that $\|X_n\| \leq Z$, then the above weak convergence is also equivalent to

$$\forall u \in \mathbb{B}(0,1)\backslash 0, \lim_{n\to+\infty} S_n(u) = S(u).$$

Perhaps the most classical univariate application of these notions is the measurement of risk, with the fundamental risk measures given by the Value-at-Risk (VaR) and the Conditional-Value-at-Risk (CVaR). The risk framework considers a vector of losses $X \in \mathbb{R}^d_+$ where each component is a positive measure. From our definitions, one can just select a point with maximal norm within a MK quantile contour (resp. superquantile) to get a multivariate VaR (resp. CVaR), see [3] for further detail. Such risk measurements summarize a given dataset to answer the following :

- With probability α, what is the worst that can happen ?
- In case the worst happens, what shall we expect, in average ?

Using MK quantiles allows to take into account the multivariate probability for the assertion "with probability α". These are illustrated in Figure 1.3, with different choices of reference measure that correspond to a center-outward or a left-to-right ordering for ν.

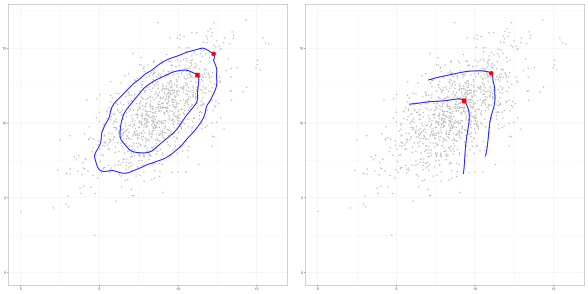

FIGURE 1.3 – *VaR and CVaR, center-outward or left-to-right ordering.*

Estimation

In practice, the estimation of $\mathbf{Q}$ amounts to solve an OT problem involving the empirical measure associated to samples $Y_1, \cdots, Y_n$ from ν. Such problem yields a quantile function whose values are constrained to belong to the (Y_i), unless one uses regularization in a second step. Interpolation is desirable as soon as the focus is on quantile contours and regions, or if it is required to compute out-of-sample estimates $\mathbf{Q}(x)$, for any x. The estimation of regularized MK quantiles is another direction of my thesis, in which we advocated the use of the entropic regularization of optimal transport, [4]. This results in the *entropic map*, that has the desirable feature of being the gradient of a convex function, even in practice, and that is rooted in the literature on computational optimal transport.

Therefore, the estimation procedure requires to solve the entropically regularized OT problem. Making use of the fact that one of the two distributions is held fixed, we were able to design a new algorithm tackling the continuous OT in the limit $n \to +\infty$, for n the size of the sample (Y_i). The idea is to parametrize dual potentials in the Kantorovich formulation of OT by their Fourier coefficients, leading to a stochastic gradient descent on absolutely summable sequences. Using stochastic algorithms has two major consequences. On the practical side, it allows to avoid the storage of the cost matrix of size n^2 between two samples of size n. On the theoretical side, one can use tools from random processes in order to obtain consistency results, as in [4] where we showed the almost sure convergence of the iterates of our algorithm.

Bibliographie

[1] Chernozhukov, M., Galichon, A., Hallin, M., Henry, M. (2017). Monge-Kantorovich depth, quantiles, ranks and signs. The Annals of Statistics, 45(1), 223-256.

[2] 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. The Annals of Statistics, 49(2), 1139-1165.

[3] Bercu, B., Bigot, J., and Thurin, G. (2023). Monge-Kantorovich superquantiles and expected shortfalls with applications to multivariate risk measurements. arXiv preprint arXiv :2307.01584.

[4] Bercu, B., Bigot, J., and Thurin, G. (2023). Stochastic optimal transport in Banach Spaces for regularized estimation of multivariate quantiles. arXiv preprint arXiv :2302.00982.