Statistics and Probability Letters 109 (2016) 232–237

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Elliptical multiple-output quantile regression and convex optimization

Marc Hallin [a,*], Miroslav Šiman [b]

[a] ECARES, Université libre de Bruxelles CP114/4, B-1050 Brussels, Belgium

[b] The Institute of Information Theory and Automation of the Czech Academy of Sciences, Pod Vodárenskou věží 4, CZ-182 08 Prague 8, Czech Republic

ARTICLE INFO

Article history:
Received 21 November 2015
Received in revised form 23 November 2015
Accepted 23 November 2015
Available online 27 November 2015

MSC:
62H12
62J99
62G05

Keywords:
Quantile regression
Elliptical quantile
Multivariate quantile
Multiple-output regression

ABSTRACT

This article extends linear quantile regression to an elliptical multiple-output regression setup. The definition of the proposed concept leads to a convex optimization problem. Its elementary properties, and the consistency of its sample counterpart, are investigated. An empirical application is provided.

1. Introduction

Due to their close relation to location and scatter, and their central role in the geometry of Gaussian and elliptical distributions, ellipsoids and the related Mahalanobis distances are quite logical tools for the statistical analysis of multivariate data. Quite naturally, thus, ellipsoids have been considered in the definition of multivariate quantiles and related concepts.

A definition of elliptical multivariate quantiles has been proposed by Hlubinka and Šiman (2013), which leads to a convex optimization problem, hence to a unique solution. That concept essentially deals with location, although its weighted version, based on covariate-driven weights, allows, in the presence of covariates, for a *local constant regression* extension. In the location case (when no covariates are available), Hlubinka and Šiman (2015) consider a more general nonlinear definition, leading to non-convex optimization. The uniqueness of the resulting quantile, therefore, is problematic.

This paper, inspired by Koenker and Bassett (1978), presents a linear multiple-output quantile regression extension of Hlubinka and Šiman (2013), and shows that it leads to a convex optimization problem with a uniquely defined solution for all multivariate continuous distributions with finite second-order moments and connected support, including those with multimodal densities that often arise in the context of mixtures (see, e.g., Došlá, 2009).

* Corresponding author.
E-mail address: mhallin@ulb.ac.be (M. Hallin).

http://dx.doi.org/10.1016/j.spl.2015.11.021
0167-7152/

Section 2 presents the new concept, Sections 3 and 4 investigate its main properties in the population case and in the sample case, and Section 5 briefly illustrates it with a real data application.

2. Definition

Let $\tau \in (0,1)$ and consider an m-dimensional response vector $\boldsymbol{Y}$ associated with a $(p+1)$-dimensional vector of regressors $(1, \boldsymbol{Z}')'$. Throughout, it is assumed that the joint distribution of $(\boldsymbol{Y}', \boldsymbol{Z}')'$ is absolutely continuous, with connected support and finite second-order moments.

In the location case (when $p=0$), Hlubinka and Šiman (2013) define the multivariate (location) elliptical τ-quantile as the ellipsoid

$$\varepsilon_\tau^{\mathrm{loc}} = \varepsilon_\tau^{\mathrm{loc}}(\boldsymbol{Y}) := \{\boldsymbol{y} \in \mathbb{R}^m : \boldsymbol{y}'\mathbb{A}_\tau \boldsymbol{y} + \boldsymbol{y}'\boldsymbol{b}_\tau - c_\tau = 0\},$$

where $\mathbb{A}_\tau \in \mathbb{R}^{m\times m}$, $\boldsymbol{b}_\tau \in \mathbb{R}^{m\times 1}$, and $c_\tau > 0$ minimize, subject to $\mathbb{A}$ being symmetric and positive semidefinite with determinant one ($\mathbb{A}$ is thus a *shape matrix* in the sense of Paindaveine (2008)), the objective function

$$\Psi_\tau^{\mathrm{loc}}(\mathbb{A}, \boldsymbol{b}, c) := \mathrm{E}\,\rho_\tau(\boldsymbol{Y}'\mathbb{A}\boldsymbol{Y} + \boldsymbol{Y}'\boldsymbol{b} - c)$$

with the usual check function $\rho_\tau(x) := x(\tau - \mathrm{I}(x<0)) = \max\{(\tau-1)x, \tau x\}$. The positive semidefiniteness of $\mathbb{A}$ and the condition on its determinant ensure that $\varepsilon_\tau^{\mathrm{loc}}$ is indeed an ellipsoid, centered at $\boldsymbol{s}_\tau := -\mathbb{A}_\tau^{-1}\boldsymbol{b}_\tau/2$, with equation $(\boldsymbol{y}-\boldsymbol{s}_\tau)'\mathbb{A}_\tau(\boldsymbol{y}-\boldsymbol{s}_\tau) = \kappa_\tau$, where $\kappa_\tau := c_\tau + \boldsymbol{b}_\tau'\mathbb{A}_\tau^{-1}\boldsymbol{b}_\tau/4$. The condition $\det(\mathbb{A}) = 1$ can be viewed as an identification constraint: for any $K>0$, the triples $(\mathbb{A}, \boldsymbol{b}, c)$ and $(K\mathbb{A}, K\boldsymbol{b}, Kc)$ indeed define the same ellipsoid.

The same definition can be reformulated as a convex optimization problem by relaxing the constraint $\det(\mathbb{A}) = 1$ into $(\det(\mathbb{A}))^{1/m} \geq 1$: the function $\mathbb{A} \mapsto (\det(\mathbb{A}))^{1/m}$, unlike $\mathbb{A} \mapsto \det(\mathbb{A})$, is concave on the cone of symmetric positive semidefinite matrices (see, e.g., Šilhavı, 2008), and the fact that $\Psi_\tau^{\mathrm{loc}}(K\mathbb{A}, K\boldsymbol{b}, Kc) = K\Psi_\tau^{\mathrm{loc}}(\mathbb{A}, \boldsymbol{b}, c)$ for any $K>0$ implies that the optimal $\mathbb{A}_\tau$ is such that $(\det(\mathbb{A}_\tau))^{1/m} = \det(\mathbb{A}_\tau) = 1$ (see Section 2 of Hlubinka and Šiman, 2013, where alternative identification constraints are also discussed).

In the presence of covariates (that is, when $p>1$), the traditional homoscedastic multiple-output linear regression model suggests, for an elliptical multiple-output regression τ-quantile, a simple equation of the form

$$(\boldsymbol{y} - \boldsymbol{\beta} - \mathbb{B}\boldsymbol{z})'\mathbb{A}_\tau(\boldsymbol{y} - \boldsymbol{\beta} - \mathbb{B}\boldsymbol{z}) - \gamma = 0$$

with some $\mathbb{A} \in \mathbb{R}^{m\times m}$, $\boldsymbol{\beta} \in \mathbb{R}^{m\times 1}$, $\mathbb{B} \in \mathbb{R}^{m\times p}$, and $\gamma > 0$. The trouble is that the corresponding objective function

$$\mathrm{E}\,\rho_\tau\big((\boldsymbol{Y} - \boldsymbol{\beta} - \mathbb{B}\boldsymbol{Z})'\mathbb{A}(\boldsymbol{Y} - \boldsymbol{\beta} - \mathbb{B}\boldsymbol{Z}) - \gamma\big)$$

is not convex in $\boldsymbol{\beta}$ and $\mathbb{B}$, so that its minimization with respect to $\mathbb{A}$, $\boldsymbol{\beta}$, $\mathbb{B}$, and γ is not a *convex* optimization problem. And the same could be said even if γ were an affine linear function of $\boldsymbol{z}$.

In order to restore convexity, consider instead the more general definition

$$\varepsilon_\tau^{\mathrm{reg}} := \{(\boldsymbol{y}', \boldsymbol{z}')' \in \mathbb{R}^{m+p} : (\boldsymbol{y} - \boldsymbol{\beta}_\tau - \mathbb{B}_\tau\boldsymbol{z})'\mathbb{A}_\tau(\boldsymbol{y} - \boldsymbol{\beta}_\tau - \mathbb{B}_\tau\boldsymbol{z}) - (\gamma_\tau + \boldsymbol{c}_\tau'\boldsymbol{z} + \boldsymbol{z}'\mathbb{C}\boldsymbol{z}) = 0\} \tag{1}$$

of an elliptical regression quantile $\varepsilon_\tau^{\mathrm{reg}} = \varepsilon_\tau^{\mathrm{reg}}(\boldsymbol{Y}, \boldsymbol{Z})$, where a quadratic form of covariate-driven scale is allowed, and $\mathbb{A}_\tau$, $\boldsymbol{\beta}_\tau$, $\mathbb{B}_\tau$, γ_τ, $\boldsymbol{c}_\tau$, and $\mathbb{C}_\tau$ jointly minimize

$$\Psi_\tau^{\mathrm{reg}} := \mathrm{E}\,\rho_\tau\big((\boldsymbol{Y} - \boldsymbol{\beta} - \mathbb{B}\boldsymbol{Z})'\mathbb{A}(\boldsymbol{Y} - \boldsymbol{\beta} - \mathbb{B}\boldsymbol{Z}) - (\gamma + \boldsymbol{c}'\boldsymbol{Z} + \boldsymbol{Z}'\mathbb{C}\boldsymbol{Z})\big)$$

under the constraint that $\mathbb{C} \in \mathbb{R}^{p\times p}$ is symmetric and $\mathbb{A} \in \mathbb{R}^{m\times m}$ is symmetric positive semidefinite with $\det(\mathbb{A}) = 1$. This minimization, however, still does not take the form of a convex optimization problem.

Let therefore $\mathbb{M} := (\mathbb{M}^1, \ldots, \mathbb{M}^6)$, with $\mathbb{M}^1 := \mathbb{A} \in \mathbb{R}^{m\times m}$ symmetric positive semidefinite, $\mathbb{M}^2 := \mathbb{B}'\mathbb{A}\mathbb{B} - \mathbb{C} \in \mathbb{R}^{p\times p}$ symmetric, $\mathbb{M}^3 := -2\mathbb{B}'\mathbb{A} \in \mathbb{R}^{p\times m}$, $\mathbb{M}^4 := -2\boldsymbol{\beta}'\mathbb{A} \in \mathbb{R}^{1\times m}$, $\mathbb{M}^5 := 2\boldsymbol{\beta}'\mathbb{A}\mathbb{B} - \boldsymbol{c}' \in \mathbb{R}^{1\times p}$, and $\mathbb{M}^6 := \boldsymbol{\beta}'\mathbb{A}\boldsymbol{\beta} - \gamma \in \mathbb{R}$. The correspondence between $\mathbb{M}$ and $(\mathbb{A}, \boldsymbol{\beta}, \mathbb{B}, \gamma, \boldsymbol{c}, \mathbb{C})$ is one-to-one, with $\mathbb{A} = \mathbb{M}^1$, $\boldsymbol{\beta} = -\frac{1}{2}\mathbb{M}^{1^{-1}}\mathbb{M}^{4'}$, $\mathbb{B} = -\frac{1}{2}\mathbb{M}^{1^{-1}}\mathbb{M}^{3'}$, $\gamma = \frac{1}{4}\mathbb{M}^4\mathbb{M}^{1^{-1}}\mathbb{M}^{4'} - \mathbb{M}^6$, $\boldsymbol{c} = \frac{1}{2}\mathbb{M}^3\mathbb{M}^{1^{-1}}\mathbb{M}^{4'} - \mathbb{M}^{5'}$, and $\mathbb{C} = \frac{1}{4}\mathbb{M}^3\mathbb{M}^{1^{-1}}\mathbb{M}^{3'} - \mathbb{M}^2$: $\mathbb{M}$ thus provides a reparametrization of the problem.

In this new parametrization, the elliptical regression quantile $\varepsilon_\tau^{\mathrm{reg}}$ can be expressed as

$$\varepsilon_\tau^{\mathrm{reg}} = \{(\boldsymbol{y}', \boldsymbol{z}')' \in \mathbb{R}^{m+p} : r(\boldsymbol{y}, \boldsymbol{z}, \mathbb{M}_\tau) = 0\}$$

where

$$\begin{aligned} r(\boldsymbol{y}, \boldsymbol{z}, \mathbb{M}) &:= \boldsymbol{y}'\mathbb{M}^1\boldsymbol{y} + \boldsymbol{z}'\mathbb{M}^2\boldsymbol{z} + \boldsymbol{z}'\mathbb{M}^3\boldsymbol{y} + \mathbb{M}^4\boldsymbol{y} + \mathbb{M}^5\boldsymbol{z} + \mathbb{M}^6 \\ &= (\boldsymbol{y} - \boldsymbol{\beta} - \mathbb{B}\boldsymbol{z})'\mathbb{A}(\boldsymbol{y} - \boldsymbol{\beta} - \mathbb{B}\boldsymbol{z}) - (\gamma + \boldsymbol{c}'\boldsymbol{z} + \boldsymbol{z}'\mathbb{C}\boldsymbol{z}), \end{aligned}$$

and $\mathbb{M}_\tau := (\mathbb{M}_\tau^1, \ldots, \mathbb{M}_\tau^6)$ jointly minimize

$$\Psi_\tau^{\mathrm{reg}} = \Psi_\tau^{\mathrm{reg}}(\mathbb{M}) := \Psi_\tau^{\mathrm{reg}}(\mathbb{M}^1, \ldots, \mathbb{M}^6) = \mathrm{E}\,\rho_\tau\big(r(\boldsymbol{Y}, \boldsymbol{Z}, \mathbb{M})\big),$$