

A semi-parametric approach to multivariate expectiles  
for outlier detection<sup>1</sup>

Oliver Lübke, Philip Kokic and Jens Breckling

WORKING PAPER SERIES  
No. 3, August 2001

QANTARIS GmbH

Hostatostraße 25  
D-65929 Höchst  
Frankfurt, Germany  
Tel.: ++49-69/3140 2311  
Fax: ++49-69/3140 2323  
Email: qantaris@freenet.de

---

<sup>1</sup>EURETIT WP 4.2: Part D of deliverables D 4.2.1 and D 4.2.2

### Abstract

In Breckling, Kocic, and Lübke (2000) a non-parametric method for multivariate  $M$ -quantile estimation was introduced. To make this method applicable to outlier detection we introduce a semi-parametric approach to multivariate expectiles that shares a special case with the above method.

**Key words:**  $M$ -quantiles, non-parametric statistics, normal distribution

## 1 Objective

Our objective is to develop a method for multivariate outlier detection. Ideally, it should be possible to relate each data point to a certain probability and direction (orientation within the whole data set). Apart from a pure ordering effect of all data points we would also like to be able – for each single data point – to define a “degree of outlyingness” by means of an objective interpretation of the probability related to this data point.

Initially the aim was to use the methods for multivariate  $M$ -quantile estimation as introduced by Breckling, Kocic, and Lübke (2000). However, it turns out that this method is not suitable for this purpose. The main problem is that the  $M$ -quantile surfaces corresponding to values of  $p$  ranging between 0 and 0.5 (the latter representing the centre of the data) are not suitable for capturing outlier behaviour of the data set: for large values of the parameter  $\delta$  ( $\delta \gtrsim 50$ ) – corresponding to a step function type of weighting scheme – and  $p = 0$  the  $M$ -quantile points are pulled towards the data points sitting right on the boundary of the convex hull of the data. Thus all these points are in this sense indistinguishable.

For small values of  $\delta$  and  $p = 0$ , on the other hand, the surfaces are still far *within* the convex hull of the data. Attempts to extend the range of  $p$  below 0 to circumvent this problem have not proven successful as there exist solutions only for “slightly negative” values of  $p$ , i.e. not “negative enough” to push the surfaces out far enough.

These  $M$ -quantile methods are strictly non-parametric. To solve the above problem we try to “add more structure” by making distributional assumptions. We then introduce a parameter that determines to which degree this distributional assumption enters our estimating equations.

## 2 The non-parametric approach

This method is described in detail in Breckling, Kocic, and Lübke (2000), here we only state the definition. For a data sample  $y_1, \dots, y_n$ ,  $y_i \in \mathbb{R}^k$ , and for given  $0 \leq p \leq 0.5$  and  $r \in \mathbb{R}^k$ ,  $\|r\| = 1$ , the multivariate  $M$ -quantile  $\theta$  is defined as the solution of the system of equations

$$\frac{1}{n} \sum_{i=1}^n (y_i - \theta) \left[ p \mathbb{1}_{\{r'(y_i - \theta) \geq 0\}} + (1 - p) \mathbb{1}_{\{r'(y_i - \theta) < 0\}} \right] w_i = 0, \quad (2.1)$$

where

$$w_i = \begin{cases} c^{-1} & \text{if } \|y_i - \theta\| < c, \\ \|y_i - \theta\|^{-1} & \text{if } \|y_i - \theta\| \geq c, \end{cases} \quad (2.2)$$

for some given  $c \geq 0$ . Note that the original definition is more general. In fact, the definition we give here is only a limiting case of the original one. However, there are two reasons why we restrict ourselves to this special case. The first is that it is the only case where the computations that follow are feasible. The second and more important reason is the following: the solution surfaces that result from  $r$  being moved around the whole  $(k - 1)$ -dimensional unit sphere extend all the way to the edge of the convex hull of the data if and only if the  $M$ -quantiles are defined as above. This is crucial as we want to be able to relate each point of the data set to (unique) values of  $p$  and  $r$ , i.e. identify each point as a solution to our estimating equations.

Since the parametric approach to be described is set in the “mean world” rather than in the “median world”, to find common grounds for the two approaches it makes sense to consider the expectile case only. This means to set  $c$  to infinity (actually, a very large value), i.e. let the multivariate expectile be the solution of the system of equations

$$\frac{1}{n} \sum_{i=1}^n (y_i - \theta) [p \mathbb{1}_{\{r'(y_i - \theta) \geq 0\}} + (1 - p) \mathbb{1}_{\{r'(y_i - \theta) < 0\}}] = 0, \quad 0 \leq p \leq \frac{1}{2}. \quad (2.3)$$

### 3 A parametric approach to multivariate expectiles

Now we assume that the data sample is perfectly normally distributed, i.e. the data points are realisations of a vector valued random variable  $Y$ , where  $Y \sim N(\mu, \Sigma)$ . The density function of  $Y$  is given by

$$f_{\mu, \Sigma}(y) = \frac{1}{\sqrt{(2\pi)^k |\det(\Sigma)|}} \exp\left(-\frac{1}{2}(y - \mu)' \Sigma^{-1} (y - \mu)\right), \quad y \in \mathbb{R}^k. \quad (3.1)$$

Under the assumption of normality, the estimating equations (2.3) can be expressed as

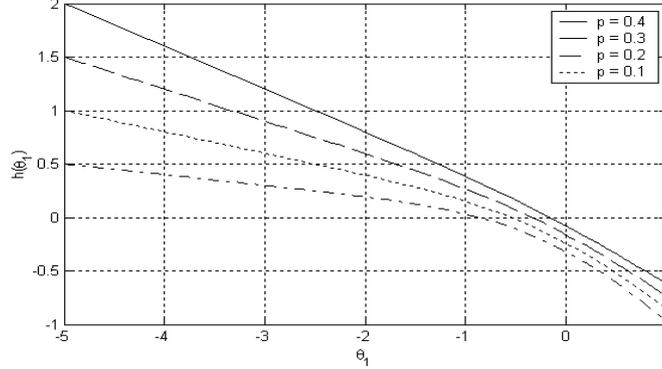
$$\int_{\mathbb{R}^k} (y - \theta) [p \mathbb{1}_{\{r'(y - \theta) \geq 0\}} + (1 - p) \mathbb{1}_{\{r'(y - \theta) < 0\}}] f_{\mu, \Sigma}(y) dy = 0, \quad (3.2)$$

where the integral is to be understood in a componentwise sense. In fact, the left hand side of (2.3) is an estimate of the left hand side of (3.2). Equation (3.2) can also be written as

$$E[g_p(Y, r, \theta)] = 0, \quad (3.3)$$

where

$$g_p(y, r, \theta) = (y - \theta) [p \mathbb{1}_{\{r'(y - \theta) \geq 0\}} + (1 - p) \mathbb{1}_{\{r'(y - \theta) < 0\}}]. \quad (3.4)$$


 Figure 1: The function  $h(\theta_1)$  for different values of  $p$ .

### 3.1 The simplified case

First we make two simplifying assumptions. Firstly we assume that  $\mu = 0$  and  $\Sigma = I$ . The second simplification is the choice of  $r = (1, 0, \dots, 0)' =: b_1$ . Then

$$\left\{ y \in \mathbb{R}^k \mid r'(y - \theta) \geq 0 \right\} = \left\{ y \in \mathbb{R}^k \mid y_1 \geq \theta_1 \right\}, \quad (3.5)$$

so for the first component of the integral we get

$$\begin{aligned} & \int_{\mathbb{R}^k} (y_1 - \theta_1) \left[ p \mathbb{1}_{\{r'(y-\theta) \geq 0\}} + (1-p) \mathbb{1}_{\{r'(y-\theta) < 0\}} \right] f_{0,I}(y) dy \\ &= (2\pi)^{-\frac{k}{2}} \int_{\mathbb{R}^k} (y_1 - \theta_1) (p \mathbb{1}_{\{y_1 \geq \theta_1\}} + (1-p) \mathbb{1}_{\{y_1 < \theta_1\}}) \exp\left(-\frac{1}{2} \sum_{j=1}^k y_j^2\right) dy \\ &= (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} (y_1 - \theta_1) (p \mathbb{1}_{\{y_1 \geq \theta_1\}} + (1-p) \mathbb{1}_{\{y_1 < \theta_1\}}) \exp\left(-\frac{1}{2} y_1^2\right) dy_1 \\ &= (2\pi)^{-\frac{1}{2}} \left[ p \left\{ (2\pi)^{\frac{1}{2}} \varphi(\theta_1) \right\} - \theta_1 (2\pi)^{\frac{1}{2}} (1 - \Phi(\theta_1)) \right] \\ &\quad + (1-p) \left\{ -(2\pi)^{\frac{1}{2}} \varphi(\theta_1) - \theta_1 (2\pi)^{\frac{1}{2}} \Phi(\theta_1) \right\} \\ &= (2p-1)(\varphi(\theta_1) + \theta_1 \Phi(\theta_1)) - p\theta_1, \end{aligned}$$

where  $\varphi$  and  $\Phi$  denote the density and cumulative distribution functions of a standard normal distribution, respectively. Since this expression depends only on the first component of  $\theta$ , the first equation of the system of equations (3.2) can be solved for  $\theta_1$  quite simply by numerical methods: set  $h(\theta_1) := (2p-1)(\varphi(\theta_1) + \theta_1 \Phi(\theta_1)) - p\theta_1$ , then  $\theta_1$  is the root of  $h$ . Figure 1 shows that  $h$  is monotonic and concave, which makes the Newton-Raphson method the natural choice for finding the unique root (pick  $\theta_1 = 0$  as an initial guess). Another point worth noticing is that the root of  $h$  approaches  $-\infty$  as  $p \downarrow 0$ .

We return to the evaluation of the integral. For  $j \geq 2$ , the  $j^{\text{th}}$  component computes as

$$\begin{aligned}
& \int_{\mathbb{R}^k} (y_j - \theta_j) [p \mathbb{1}_{\{r'(y-\theta) \geq 0\}} + (1-p) \mathbb{1}_{\{r'(y-\theta) < 0\}}] f_{0,I}(y) dy \\
&= (2\pi)^{-\frac{k}{2}} \int_{\mathbb{R}^k} (y_j - \theta_j) (p \mathbb{1}_{\{y_1 \geq \theta_1\}} + (1-p) \mathbb{1}_{\{y_1 < \theta_1\}}) \exp\left(-\frac{1}{2} \sum_{j=1}^k y_j^2\right) dy \\
&= (2\pi)^{-1} \left[-\theta_j (2\pi)^{\frac{1}{2}}\right] \left[p(1 - \Phi(\theta_1)) (2\pi)^{\frac{1}{2}} + (1-p) \Phi(\theta_1) (2\pi)^{\frac{1}{2}}\right] \\
&= \theta_j [(2p-1)\Phi(\theta_1) - p].
\end{aligned}$$

Because  $0 \leq p \leq \frac{1}{2}$ , the expression in brackets in the last line above is strictly negative, so the solution to (3.2) is

$$\theta_j = 0, \quad 2 \leq j \leq k. \quad (3.6)$$

**Remark 3.1.** Recall that we chose  $r = b_1$ , so  $\theta$  turns out to be a multiple of  $r$ . This result makes intuitive sense. Note, however, that this does not hold true in the non-parametric approach.

### 3.2 The general case

Now we show how to proceed in the general case. Clearly, to assume a mean of zero is no restriction. Now suppose that  $Y$  is  $N(0, \Sigma)$ -distributed. Because  $\Sigma$  is symmetric there is an orthogonal transformation that turns  $\Sigma$  into a diagonal matrix. If this transformation is applied to the data set it is only rotated as a whole without changing the geometric qualities. Thus it is justified to assume that  $\Sigma$  is diagonal in the first place.

Let  $r$  be an arbitrary vector of length 1. The idea is to transform  $Y$  and  $r$  in such a way that the estimating equation (3.3) can be related to the simpler case from the previous section. Let  $\Sigma^{\frac{1}{2}}$  denote the matrix for which  $\Sigma^{\frac{1}{2}} \Sigma^{\frac{1}{2}} = \Sigma$ , and  $\Sigma^{-\frac{1}{2}}$  the inverse of  $\Sigma^{\frac{1}{2}}$ . Recall the following two simple results.

**Lemma 3.2.**

1. If  $Y \sim N(0, \Sigma)$  with  $\Sigma$  diagonal, then  $\Sigma^{-\frac{1}{2}} Y \sim N(0, I)$ .
2. If  $T$  is an orthogonal transformation and  $Y \sim N(0, I)$ , then also  $TY \sim N(0, I)$ .

*Proof.*

1.  $\text{Cov}[\Sigma^{-\frac{1}{2}} Y] = E[\Sigma^{-\frac{1}{2}} Y Y' \Sigma^{-\frac{1}{2}}] = \Sigma^{-\frac{1}{2}} \Sigma \Sigma^{-\frac{1}{2}} = I$ , as  $\Sigma$  is diagonal.
2.  $\text{Cov}[TY] = E[T Y Y' T'] = T I T' = I$ . □

Now set  $\tilde{Y} := \Sigma^{-\frac{1}{2}}Y$ ,  $\tilde{r} := \Sigma^{\frac{1}{2}}r/\|\Sigma^{\frac{1}{2}}r\|$ . Let  $T$  denote an orthogonal transformation (rotation) that maps  $\tilde{r}$  to  $b_1$  and define  $\hat{Y} := T\tilde{Y}$ ,  $\hat{r} := T\tilde{r}$ .<sup>1</sup> Because of lemma 3.2 we know that  $\hat{Y} \sim N(0, I)$ . Thus we can use the methods described in the previous section to find a solution  $\hat{\theta}$  to the equation

$$E[g_p(\hat{Y}, \hat{r}, \hat{\theta})] = 0. \quad (3.7)$$

From this we can construct a solution for the general case in the following way.

**Theorem 3.3.** *Let  $\hat{\theta}$  be a solution to (3.7). Then the vector  $\theta := \Sigma^{\frac{1}{2}}T'\hat{\theta}$  solves the estimating equation (3.3).*

*Proof.* First note that  $Y = \Sigma^{\frac{1}{2}}T'\hat{Y}$ ,  $r = \|\Sigma^{\frac{1}{2}}r\|\Sigma^{-\frac{1}{2}}T'\hat{r}$ , and

$$r'(Y - \theta) = \|\Sigma^{\frac{1}{2}}r\|(\Sigma^{-\frac{1}{2}}T'\hat{r})'\Sigma^{\frac{1}{2}}T'(\hat{Y} - \hat{\theta}) = \|\Sigma^{\frac{1}{2}}r\|\hat{r}'(\hat{Y} - \hat{\theta}). \quad (3.8)$$

Equation (3.8) means that

$$r'(Y - \theta) \geq 0 \iff \hat{r}'(\hat{Y} - \hat{\theta}) \geq 0. \quad (3.9)$$

Hence,

$$\begin{aligned} E[g_p(Y, r, \theta)] &= E[g_p(\Sigma^{\frac{1}{2}}T'\hat{Y}, \|\Sigma^{\frac{1}{2}}r\|\Sigma^{-\frac{1}{2}}T'\hat{r}, \Sigma^{\frac{1}{2}}T'\hat{\theta})] \\ &= E[\Sigma^{\frac{1}{2}}T'(\hat{Y} - \hat{\theta})\{p\mathbb{1}_{\{\hat{r}'(\hat{Y} - \hat{\theta}) \geq 0\}} + (1-p)\mathbb{1}_{\{\hat{r}'(\hat{Y} - \hat{\theta}) < 0\}}\}] \\ &= \Sigma^{\frac{1}{2}}T'E[g_p(\hat{Y}, \hat{r}, \hat{\theta})] \\ &= 0. \end{aligned} \quad \square$$

An immediate consequence from the above considerations is the following result.

**Theorem 3.4.** *Let  $Y \sim N(0, \Sigma)$  and  $r \in \mathbb{R}^k$  with  $\|r\| = 1$ . Let  $\theta$  be the corresponding solution of (3.3). Then  $\theta$  is a multiple of  $\Sigma r$ .*

*Proof.* With the notation from above, let  $\hat{\theta}$  be the solution of (3.3) corresponding to  $\hat{Y}$  and  $\hat{r}$ . Since  $\hat{Y} \sim N(0, I)$  and  $\hat{r} = b_1$ , from remark 3.1 we know that  $\hat{\theta}$  is a multiple of  $\hat{r}$ , say  $\hat{\theta} = \xi\hat{r}$ . With theorem 3.3 it follows that

$$\theta = \Sigma^{\frac{1}{2}}T'\hat{\theta} = \Sigma^{\frac{1}{2}}T'\xi\hat{r} = (\xi/\|\Sigma^{\frac{1}{2}}r\|)\Sigma^{\frac{1}{2}}T'T\Sigma^{\frac{1}{2}}r = (\xi/\|\Sigma^{\frac{1}{2}}r\|)\Sigma r. \quad \square$$

Before we outline how the parametric and the non-parametric approaches can be mixed we discuss how the solution surfaces in the parametric approach can be interpreted in terms of probability.

<sup>1</sup>This transformation  $T$  is not unique. However, all that is important is that it is orthogonal to ensure that it does not change the geometry of the data set when applied to it. A possibility to find such a transformation  $T$  is described in section A.1.

## 4 Probabilistic interpretation of the solution surfaces

So far there is no objective interpretation of the meaning of the parameter  $p$ . To achieve such an interpretation we proceed in two steps. First we introduce a parametric notion of multivariate quantiles (rather than expectiles) that allows for a direct probabilistic interpretation.<sup>2</sup> In a second step we show that this approach is in fact equivalent (in a certain sense) to the parametric expectile approach introduced in the previous section.

### 4.1 A parametric approach to multivariate quantiles

Let  $Y$  be a normally distributed random variable that takes values in  $\mathbb{R}^k$ , more precisely  $Y \sim N(0, \Sigma)$ . The idea is to define the quantile surfaces as the iso-height surfaces of the density function of  $Y$ : for a given value  $0 \leq \bar{p} \leq 1$ , we define the corresponding quantile surface as the particular iso-weight surface such that the region enclosed by this surface has probability mass  $\bar{p}$ .

#### 4.1.1 The simplified case

For a start we assume that  $\Sigma = I$ . Since in this case the iso-weight surfaces are spherical we only have to find the radius  $\varrho$  such that the probability mass within such a sphere is equal to some given value  $0 \leq \bar{p} \leq 1$ . Let this sphere be denoted by  $S_\varrho^{k-1}$ . To compute  $\varrho$  we have to solve the equation

$$\int_{S_\varrho^{k-1}} f_{0,I}(y) dy = \bar{p}, \quad 0 \leq \bar{p} \leq 1. \quad (4.1)$$

Because of the above assumption the random variable  $Y$  can be interpreted as a set of  $k$  independent, standard normally distributed random variables  $Y_1, \dots, Y_k$ . Thus the integral in (4.1) can be written as

$$\int_{S_\varrho^{k-1}} f_{0,I}(y) dy = P(Y_1^2 + \dots + Y_k^2 \leq \varrho^2) = F_{\chi^2, k}(\varrho^2), \quad (4.2)$$

where  $F_{\chi^2, k}$  denotes the cumulative distribution function of a  $\chi^2$ -distribution with  $k$  degrees of freedom. To get in correspondence with the setup for the non-parametric expectiles as defined in section 2 we have to perform the transformation  $\bar{p} \mapsto \frac{1}{2}(1 - \bar{p})$  such that  $0 \leq \bar{p} \leq 0.5$ , where  $\bar{p} = 0.5$  corresponds to the mean and  $\bar{p} = 0$  to infinity. With this the solution to (4.1) is given by

$$\varrho^2 = F_{\chi^2, k}^{-1}(1 - 2\bar{p}), \quad 0 \leq \bar{p} \leq \frac{1}{2}. \quad (4.3)$$

---

<sup>2</sup>For convenience, this approach will henceforth be referred to as the *(parametric) quantile approach*, whereas the approach from section 3 will be referred to as the *(parametric) expectile approach*. Note, however, that the two approaches differ by more than just the different notions of quantile and expectile.

This yields a sphere with radius  $\varrho$  as the  $(k - 1)$ -dimensional quantile surface. In correspondence to the other approaches we want to define a multivariate quantile with respect to a given directional unit vector  $\bar{r}$ . A straight forward way of achieving this is to define the multivariate parametric quantile  $\bar{\theta}_{\bar{p}, \bar{r}}$  as the point

$$\bar{\theta}_{\bar{p}, \bar{r}} = -\varrho \bar{r}. \quad (4.4)$$

The minus sign is due to the interpretation of  $\bar{p}$ : since  $\bar{p} = \frac{1}{2}$  is supposed to represent the centre of the data, values of  $\bar{p}$  less than  $\frac{1}{2}$  should lie in the direction opposite to that of  $\bar{r}$ . This interpretation is analogue to the one in the approaches above. Note, however, that  $r$  and  $\bar{r}$  are not entirely equivalent. In (4.4), the quantile lies by definition on the line given by  $\bar{r}$ . In the non-parametric approach this is not necessarily true.

#### 4.1.2 The general case

Suppose that  $\Sigma$  is diagonal with diagonal entries  $\sigma_1^2, \dots, \sigma_k^2$ . The first thing to notice is that the iso-weight surfaces of  $Y$  are now  $(k - 1)$ -dimensional ellipsoids  $E_{\varrho, \Sigma}^{k-1}$  given by the equation

$$\sum_{j=1}^k \frac{y_j^2}{\sigma_j^2} = \varrho^2, \quad (4.5)$$

as can easily be seen by setting  $f_{0, \Sigma}$  equal to a constant. Thus it is our aim to find some  $\varrho > 0$  such that

$$\int_{E_{\varrho, \Sigma}^{k-1}} f_{0, \Sigma}(y) dy = 1 - 2\bar{p}, \quad 0 \leq \bar{p} \leq \frac{1}{2}. \quad (4.6)$$

Now consider the transformation  $T: \mathbb{R}^k \rightarrow \mathbb{R}^k$ ,  $y_j \mapsto \frac{y_j}{\sigma_j}$ . Then  $\det(T') = \prod_{j=1}^k \sigma_j^{-1}$  and  $T(E_{\varrho, \Sigma}^{k-1}) = S_{\varrho}^{k-1}$ , so

$$\int_{E_{\varrho, \Sigma}^{k-1}} f_{0, \Sigma}(y) dy = \int_{E_{\varrho, \Sigma}^{k-1}} (f_{0, I} \circ T)(y) |\det(T')| dy = \int_{S_{\varrho}^{k-1}} f_{0, I}(y) dy. \quad (4.7)$$

Thus it is again equation (4.3) that lets us compute  $\varrho$ . To get the multivariate parametric quantile  $\bar{\theta}_{\bar{p}, \bar{r}}$  for given  $\bar{p}$  and  $\bar{r}$  it is natural to define  $\bar{\theta}_{\bar{p}, \bar{r}}$  as the intersect of the line given by  $\bar{r}$  and the ellipsoid  $E_{\varrho, \Sigma}^{k-1}$ . For this we have to find  $\lambda$  such that

$$\sum_{j=1}^k \frac{(\lambda r_j)^2}{\sigma_j^2} = \varrho^2, \quad \text{i.e.} \quad \lambda = \pm \varrho \left( \sum_{j=1}^k \frac{\bar{r}_j^2}{\sigma_j^2} \right)^{-\frac{1}{2}}. \quad (4.8)$$

We summarise this result in the following definition.

**Definition 4.1.** Let  $Y \sim N(0, \Sigma)$ , where  $\Sigma$  is a diagonal matrix with diagonal elements  $\sigma_1^2, \dots, \sigma_k^2$ . Further let  $\bar{r}$  be a directional unit vector and  $0 \leq \bar{p} \leq \frac{1}{2}$ . We define the  $k$ -dimensional parametric quantile  $\bar{\theta}_{\bar{p}, \bar{r}}$  as the point

$$\bar{\theta}_{\bar{p}, \bar{r}} = \lambda \bar{r}, \quad \text{where} \quad \lambda = -\left(F_{\chi^2, k}^{-1}(1 - 2\bar{p})\right)^{\frac{1}{2}} \left(\sum_{j=1}^k \frac{\bar{r}_j^2}{\sigma_j^2}\right)^{-\frac{1}{2}}. \quad (4.9)$$

## 4.2 The relationship between the solution surfaces

By construction, the solution surfaces to the estimating equations (4.6) in the parametric quantile approach are ellipsoids  $E_{\varrho, \Sigma}^{k-1}$  defined by

$$\sum_{j=1}^k \frac{x_j^2}{\sigma_j^2} = \varrho^2. \quad (4.10)$$

In this section we show that the solution surfaces to the estimating equations (3.3) in the parametric expectile approach – as  $r$  is moved around the unit sphere  $S^{k-1}$  – are ellipsoids of the exact same shape, except that to a given solution surface there correspond different values of  $p$  and  $\bar{p}$  in the two approaches.

### 4.2.1 The simplified case

We start by considering the case  $Y \sim N(0, I)$  again. As shown, the solution surfaces in the quantile approach are spherical. The same holds true for the solution surfaces of (3.3) in the expectile approach:

**Theorem 4.2.** *If  $\Sigma = I$ , then the solution surfaces of (3.3) are spherical.*

*Proof.* Let  $r$  be an arbitrary unit vector, and let  $\theta$  such that  $E[g(Y, r, \theta)] = 0$ . Let  $\hat{r}$  be another arbitrary unit vector and let  $T$  denote an orthogonal transformation that rotates  $r$  into  $\hat{r}$ . A similar computation as above and applying lemma 3.2 shows that for  $\hat{\theta} := T\theta$  we have

$$E[g_p(Y, \hat{r}, \hat{\theta})] = E[g_p(TY, Tr, T\theta)] = TE[g_p(Y, r, \theta)] = 0, \quad (4.11)$$

so  $\hat{\theta}$  is a solution to (3.3). Because  $T$  is orthogonal we know that  $\|\theta\| = \|\hat{\theta}\|$ . This completes the proof.  $\square$

### 4.2.2 The general case

Now define  $\tilde{Y} := \Sigma^{\frac{1}{2}}Y$ , so  $\tilde{Y} \sim N(0, \Sigma)$ . By the same arguments as above we see that for  $\tilde{r} := \Sigma^{-\frac{1}{2}}r / \|\Sigma^{-\frac{1}{2}}r\|$  and  $\tilde{\theta} := \Sigma^{\frac{1}{2}}\theta$  we have

$$E[g_p(\tilde{Y}, \tilde{r}, \tilde{\theta})] = 0, \quad (4.12)$$

so  $\tilde{\theta}$  is a solution of the estimating equation. In other words, if  $\Theta \subset \mathbb{R}^k$  denotes the solution surface corresponding to  $Y \sim N(0, I)$  as  $r$  is moved around the  $(k-1)$ -dimensional unit sphere, then  $\tilde{\Theta} := \Sigma^{\frac{1}{2}}\Theta$  is the solution surface corresponding to  $\tilde{Y} \sim N(0, \Sigma)$ . Let  $\theta \in \Theta$  and define  $\tilde{\theta} := \Sigma^{\frac{1}{2}}\theta \in \tilde{\Theta}$ . Then, because  $\Theta$  is spherical,

$$\sum_{j=1}^k \frac{\tilde{\theta}_j^2}{\sigma_j^2} = \sum_{j=1}^k \frac{(\sigma_j \theta_j)^2}{\sigma_j^2} = \sum_{j=1}^k \theta_j^2 = \varrho^2, \quad (4.13)$$

where  $\varrho$  is the radius of  $\Theta$ . Thus  $\tilde{\Theta}$  in fact coincides with the ellipsoid  $E_{\varrho, \Sigma}^{k-1}$ , which is a simple expansion or contraction of the solution surface we get in the purely parametric approach. We summarise this result in the following theorem.

**Theorem 4.3.** *If  $Y \sim N(0, \Sigma)$ ,  $\Sigma$  diagonal, then the solution surface of the estimating equation (3.3) coincides with the ellipsoid  $E_{\varrho, \Sigma}^{k-1}$ , where  $\varrho$  is the radius of the spherical solution surface corresponding to  $Y \sim N(0, I)$ .*

#### 4.2.3 The precise correspondence of the surfaces

Obviously, for  $Y \sim N(0, \Sigma)$  and a given value  $\varrho > 0$ , the ellipsoid  $E_{\varrho, \Sigma}^{k-1}$  is the solution surface for the estimating equation (4.6) in the quantile approach for some value  $\bar{p}$ . On the other hand,  $E_{\varrho, \Sigma}^{k-1}$  is also the solution surface for the estimating equation (3.3) in the expectile approach for some value  $p$ . What is the precise relationship between  $\bar{p}$  and  $p$ ?

According to theorem 4.3,  $\varrho$  equals the radius of the spherical solution surface that results from the estimating equation (3.3) for a standard normally distributed random variable. As shown in section 3.1, for  $r = b_1$  the solution  $\theta$  is given by the system of equations

$$(2p-1)(\varphi(\theta_1) + \theta_1 \Phi(\theta_1)) - p\theta_1 = 0 \quad (4.14)$$

$$\theta_j = 0, \quad 2 \leq j \leq k. \quad (4.15)$$

This shows that  $\varrho$  has to equal the absolute value of  $\theta_1$ . For  $r = b_1$  we know that  $\theta_1$  has to be negative, so in fact

$$\varrho = -\theta_1. \quad (4.16)$$

According to the quantile approach, on the other hand, the relationship between  $\varrho$  and  $\bar{p}$  is given by equation (4.3), i.e.

$$\varrho = \sqrt{F_{\chi^2, k}^{-1}(1-2\bar{p})}. \quad (4.17)$$

Combining these two identities and plugging that into (4.14) yields the following result.

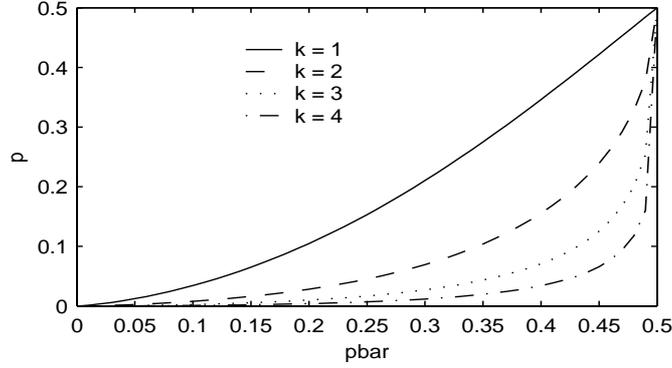


Figure 2: The value of  $p$  as a function of  $\bar{p}$  for different values of  $k$ .

**Theorem 4.4.** For given  $\bar{p}$  let  $\bar{\Theta}$  denote the solution surface in the parametric quantile approach. For given  $p$  let  $\Theta$  denote the solution surface corresponding to the estimating equation (3.3) in the parametric expectile approach. Then  $\Theta = \bar{\Theta}$  if and only if

$$p = \frac{\varphi(\varrho) - \varrho\Phi(-\varrho)}{2(\varphi(\varrho) - \varrho\Phi(-\varrho)) + \varrho}, \quad \varrho = \sqrt{F_{\chi^2, k}^{-1}(1 - 2\bar{p})}. \quad (4.18)$$

This relationship is plotted in figure 2 for different values of  $k$ . It allows us to interpret the solution surfaces in the expectile approach in the same probabilistic manner as in the quantile approach. Equation (4.18) can also be interpreted as a function in  $\bar{p}$  (of  $p$ ) to obtain the inverse relationship. The necessary computations for the numerical solution can be found in the appendix.

We conclude this section by combining theorems 4.4 and 3.4 to establish a (1:1)-correspondence between the expectile and the quantile approach.

**Theorem 4.5.** Let  $Y \sim N(0, \Sigma)$ ,  $r, \bar{r} \in \mathbb{R}^k$  with  $\|r\| = \|\bar{r}\| = 1$ , and  $0 \leq p, \bar{p} \leq \frac{1}{2}$ . Let  $\theta_{p,r}$  be the solution to (3.3) and  $\bar{\theta}_{\bar{p},\bar{r}}$  the solution in the quantile approach, i.e. let  $\bar{\theta}_{\bar{p},\bar{r}}$  be given by (4.9). Then  $\theta_{p,r} = \bar{\theta}_{\bar{p},\bar{r}}$  if and only if  $p$  is given by (4.18) and  $r = \Sigma^{-1}\bar{r}/\|\Sigma^{-1}\bar{r}\|$ .

## 5 The semi-parametric approach

### 5.1 Definition

Our original aim was to find a mixture of the non-parametric and the parametric approach to multivariate expectiles. Recall that we are dealing with a set of data points  $y_1, \dots, y_n$ . Assume that the data set has an estimated mean of 0 and an estimated covariance matrix  $\Sigma$ ,  $\Sigma$  diagonal.<sup>3</sup> Then we propose the following definition.

<sup>3</sup>We emphasise again that these assumptions are no restrictions as any data set can be transformed – without change of geometric properties – to fulfil these conditions.

**Definition 5.1.** Let  $0 \leq p \leq \frac{1}{2}$  and  $r \in \mathbb{R}^k$  with  $\|r\| = 1$ , and let  $0 \leq \eta \leq 1$ . The *semi-parametric expectile*  $\theta = \theta_{p,r,\eta}$  is defined as the solution to the system of equations

$$\eta \frac{1}{n} \sum_{i=1}^n g_p(y_i, r, \theta) + (1 - \eta) \int_{\mathbb{R}^k} g_p(y, r, \theta) f_{0,\Sigma}(y) dy = 0, \quad (5.1)$$

where  $g_p$  is given by (3.4).

The parameter  $\eta$  determines to which degree each of the two approaches enters the estimating equations. The way the two approaches are mixed may seem a bit arbitrary at first glance. However, there are arguments to justify this procedure.

Firstly, as mentioned before, the sum in the estimating equations above is an estimate of the integral, so we are mixing objects of the same kind. Secondly, equation (5.1) could be viewed as a penalty type of estimating equation: if one assumes an underlying normal distribution, the sum in (5.1) can be interpreted as a penalty term compensating for departure from normality.

Note that in definition 5.1 the density of the normal distribution could in principle be replaced by any other density function. In most cases, however, the resulting estimating equations will hardly be analytically or numerically feasible. Also, the theory concerning the probabilistic interpretation we developed in section 4 is valid only for the normal distribution.

## 5.2 Reformulation of the estimating equations

To get the estimating equations (5.1) in a numerically tractable form we prove a simple result for the non-parametric expectile approach that is exactly analogue to theorem 3.3 in the parametric expectile approach.

**Theorem 5.2.** Let  $0 \leq p \leq \frac{1}{2}$ ,  $r \in \mathbb{R}^k$  with  $\|r\| = 1$ , and  $\Sigma \in \mathbb{R}^{k \times k}$  a diagonal matrix. Let  $\hat{y}_i$ ,  $\hat{r}$  and  $T$  be defined as in theorem 3.3. Let  $\hat{\theta}_{p,\hat{r}}$  be the solution of (2.3) with respect to  $\hat{r}$ . Then  $\theta_{p,r} := \Sigma^{\frac{1}{2}} T' \hat{\theta}$  is the solution to (2.3) with respect to  $r$ .

*Proof.* The estimating equation (2.3) can be written as

$$\frac{1}{n} \sum_{i=1}^n g_p(y_i, r, \theta) = 0, \quad (5.2)$$

and, as before,  $r'(y_i - \theta) \geq 0$  if and only if  $\hat{r}'(\hat{y}_i - \hat{\theta}) \geq 0$ . Thus  $g_p(y_i, r, \theta) = \Sigma^{\frac{1}{2}} T' g_p(\hat{y}_i, \hat{r}, \hat{\theta})$ . This completes the proof.  $\square$

With theorems 3.3 and 5.2 we can restrict ourselves to the case where  $\Sigma = I$  and  $r = b_1$ , because (5.1) is a linear combination of the estimating equations of the two approaches. Note, however, that we have to redo the transformation of  $r$  to  $b_1$  for each new  $r$ , because the transformation of the data points  $y_i$  depends specifically on this transformation. This is in contrast to the parametric approach where the random variable  $Y$  ends up being standard normally distributed in any case.

Nonetheless theorem 5.2 is very helpful because now we can obtain an integral free formulation of the semi-parametric estimating equations (5.1) by applying the results from section 3.1: By setting

$$M_1 := \{i | y_{i1} \geq \theta_1\}, \quad M_2 := \{i | y_{i1} < \theta_1\}, \quad (5.3)$$

the system of equations (5.1) can be written as

$$\begin{aligned} \eta \frac{1}{n} \left( p \sum_{M_1} (y_{i1} - \theta_1) + (1-p) \sum_{M_2} (y_{i1} - \theta_1) \right) \\ + (1-\eta) \left( (2p-1) [\varphi(\theta_1) + \theta_1 \Phi(\theta_1)] - p\theta_1 \right) = 0, \end{aligned} \quad (5.4)$$

$$\begin{aligned} \eta \frac{1}{n} \left( p \sum_{M_1} (y_{ij} - \theta_j) + (1-p) \sum_{M_2} (y_{ij} - \theta_j) \right) \\ + (1-\eta) \theta_j \left( (2p-1) \Phi(\theta_1) - p \right) = 0, \quad 2 \leq j \leq k. \end{aligned} \quad (5.5)$$

The first of these equations depends only on  $\theta_1$  and can thus be solved quite simply by numerical methods. In fact it is easy to see that the function of which we are trying to find a root is continuous, piece-wise affine linear, strictly concave and decreasing, so the Newton-Raphson method with starting value  $\theta_1 = 0$  is the best choice.<sup>4</sup> This computed value can then be plugged into (5.5). What remains is a linear equation in  $\theta_j$ . Collecting terms yields

$$\theta_j = \frac{\eta \frac{1}{n} \left( p \sum_{M_1} y_{ij} + (1-p) \sum_{M_2} y_{ij} \right)}{\eta \frac{1}{n} \left( p \sum_{M_1} 1 + (1-p) \sum_{M_2} 1 \right) - (1-\eta) \left( (2p-1) \Phi(\theta_1) - p \right)}, \quad 2 \leq j \leq k. \quad (5.6)$$

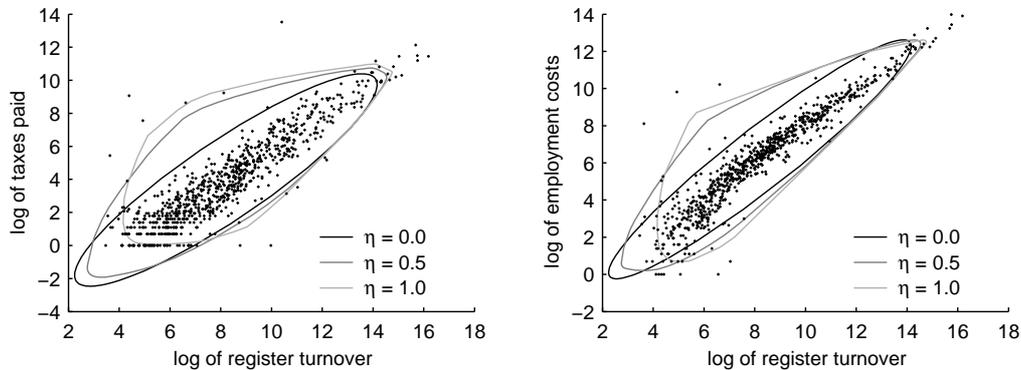
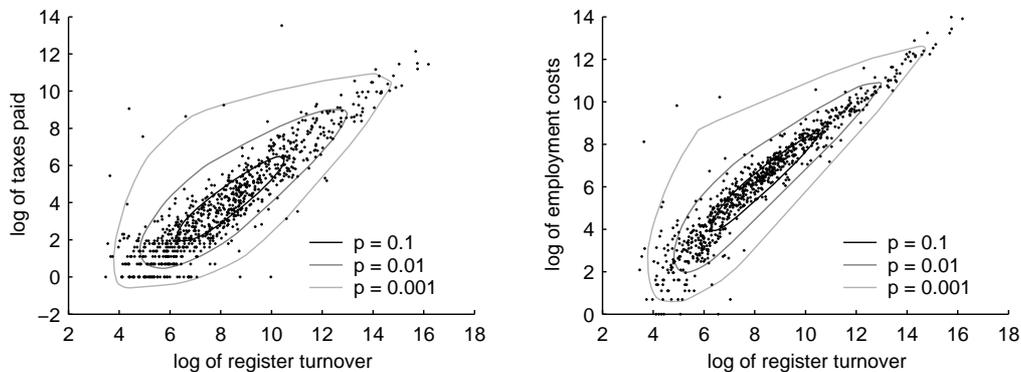
This means that the system of estimating equations (5.1) can basically be reduced to a univariate root-finding problem (and the necessary pre-transformations). Algorithms for the estimation of the semi-parametric expectiles, and the inverse problem of finding  $p$  and  $r$  for a given point in the data set are presented in the appendix.

## 6 Application

To illustrate the technique described in section 5 it was applied to data from the Annual Business Inquiry (ABI), a large scale survey run by the UK Office for National Statistics (ONS). In particular, our interest focused on the choice of the mixing parameter  $\eta$ .

A range of business and balance sheet information is collected in the ABI. It is a broad business establishment survey covering most industries in UK, except for agriculture. The data made available for this study by ONS consists of 6 study variables: turnover, employment costs, purchases of goods and services, taxes paid,

<sup>4</sup>An immediate consequence of the monotony is the uniqueness of the solution!

Figure 3: Expectile curves for different values of  $\eta$ ,  $p = 0.001$ .Figure 4: Expectile curves for different values of  $p$ ,  $\eta = 0.9$ .

cost of all capital assets required, and proceeds from capital asset disposals. In addition there are several auxiliary variables included with the data which have been extracted from the UK business register: employment, turnover, and class of activity. These variables are often used by ONS as covariates for estimation of means and totals as their values are known for all establishments in the target population: - in particular they are used to construct the survey weight, which was also included in the study data set.

Typically the study variables have a significant proportion of zeros (the remaining values being positive) and as modelling this kind of population added considerable complexity to our task, we decided to delete all observations where one or more of the study variables was zero. After deletion there remained a total of 788 observations; and subsequently there seemed to be little benefit in incorporating the survey weight into the analysis, so this was also ignored. As it is easy to illustrate, we decided to only consider the 2-dimensional case, and confined our attention to the variables register turnover, employment costs and taxes paid, although the method is clearly applicable to higher dimensional data.

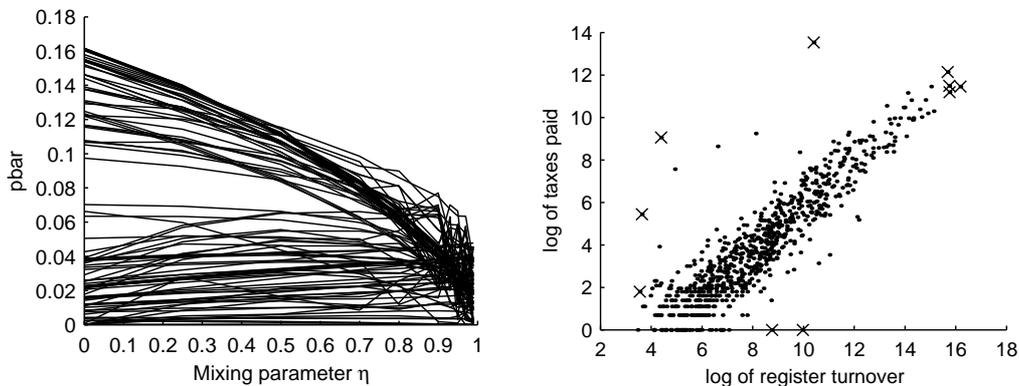


Figure 5: ‘Ridge’ plot (left) and corresponding outlier plot (right). In the left plot only observations with  $\bar{p} < 0.05$  when  $\eta = 1$  are shown. In the right plot outliers are those points with  $\bar{p} < 0.01$  when  $\eta = 0.9$ .

Figure 3 illustrates the logged data and expectile curves for several different values of  $\eta$  and a small value of  $p$ , while figure 4 shows some of the extreme curves for  $\eta$  close to 1 (close to nonparametric expectiles). Clearly it can be seen that the extreme expectile curve is quite dependent on  $\eta$  and that the departure from bivariate normality in this case is quite significant. Indeed, in the right-hand plots one can see some evidence of non-linearity, and in all plots there is strong evidence of heteroscedasticity. The expectile curves for small values of  $\eta$  capture the general shape of the data set quite well, although non-linearity is not well described. These facts are well known from earlier work on nonparametric  $M$ -quantiles, see Breckling, Kopic, and Lübke (2000).

As noted in the introduction, the fundamental problem with purely nonparametric expectiles is that they will assign the value  $p = 0$  to all points on the convex hull of the data and, as can be seen in the example presented here, this result is inappropriate. One would expect that as  $\eta$  moved away from 1 the  $p$ -values for many points on or near the convex hull of the data would increase rapidly, while those points that were indeed outlying would continue to have small values of  $p$ . On the other hand for  $p$  close to zero some false outliers may be detected because of the invalidity of the normal assumption. As shown on the left-hand plot in figure 5, this is indeed the case. One can interpret this plot in a somewhat similar fashion to a ridge plot used in ridge regression. In this case, however, one should try and choose  $\eta$  as close as possible to 1 in order to minimise the degree of bias associated with departures from normality of the underlying data, but at the same time choose  $\eta$  sufficiently far from 1 in order to avoid the problem noted above. A suitable compromise in this case seems to be  $\eta = 0.9$ . The right-hand plot shows the outliers identified by this procedure when using this particular choice of  $\eta$ . Here the cutoff value  $\bar{p} < 0.01$  was chosen arbitrarily.

## References

Breckling, J., P. Kopic, and O. Lübke (2000). A New Definition of Multivariate  $M$ -quantiles Based on a Generalisation of the Univariate Estimating Equations. Working paper 1, Insiders Financial Solutions GmbH, Mainz.

## A Description of the algorithms

So far we implicitly restricted ourselves to the problem of finding the expectile  $\theta$  for given values of  $p$  and  $r$ . For the task of outlier detection it is the reverse problem that is of interest: For a given element of the data set, find the corresponding values of  $p$  and  $r$ .<sup>5</sup> For both of these cases we shortly summarise the necessary algorithms which have already been implemented. Initial tests indicate that they both work fine.

### A.1 An application of the $QR$ -decomposition

At certain spots in the algorithms we have to find an orthogonal transformation  $T$  that maps a unit vector  $r$  to the vector  $b_1 = (1, 0, \dots, 0)'$ . One possibility to construct such a transformation is by a so-called  $QR$ -decomposition: the vector  $r$  can be written as

$$r = QR, \quad (\text{A.1})$$

where  $Q \in \mathbb{R}^{k \times k}$  is an orthogonal matrix and  $R \in \mathbb{R}^k$  is an upper triangular “matrix” – which in this case means that  $R = (R_1, 0, \dots, 0)'$ . Multiplying (A.1) by  $Q'$  from left yields  $Q'r = R$ . Thus, because  $\|r\| = 1$  and  $Q$  is orthogonal, we know that

$$1 = \|Q'r\| = \|R\| = |R_1|, \quad (\text{A.2})$$

so  $R_1 \in \{-1, 1\}$ . This means that  $R \in \{-b_1, b_1\}$ . Hence, setting  $T := R_1 Q'$  yields

$$Tr = b_1, \quad (\text{A.3})$$

where  $T$  is orthogonal. We perform the  $QR$ -decomposition of  $r$  with the Matlab-function `qr`.

### A.2 The expectile finding problem

1. Choose parameters:
  - (a) Choose  $0 \leq \eta \leq 1$ .
  - (b) Choose  $0 \leq p \leq \frac{1}{2}$ .
  - (c) Choose directional vector.
2. Perform pre-transformations:

---

<sup>5</sup>For ordering purposes only the value of  $p$  is needed.

- (a) If wanted, transform  $p$  according to equation (4.18).
  - (b) Shift data set to get a mean of zero.
  - (c) Orthogonally transform it to make the covariance matrix diagonal; call this new covariance matrix  $\Sigma$ . Apply the same transformation to the directional vector and call this  $r$ .
  - (d) Transform the data set by  $\Sigma^{-\frac{1}{2}}$  and  $r$  by  $\Sigma^{\frac{1}{2}}$  to get  $\tilde{r}$ .
  - (e) Transform the normed version of  $\tilde{r}$  by some orthogonal  $T$  to map it to  $b_1$  (see section A.1). Apply the same transformation to the data set.
3. Solve the estimating equations:
    - (a) Solve equation (5.4) for  $\theta_1$  by Newton-Raphson with starting value 0.
    - (b) Plug this value into equation (5.6) to get  $\theta_j$ ,  $2 \leq j \leq k$ .
  4. Apply the pre-transformations backwards to get the final value for  $\theta$ .

### A.3 The reverse problem

Here the main idea is to use an iterative procedure applying the estimating equations from the “forward problem”. This algorithm is more or less independent of the precise nature of the definition of the expectile. We only assume that the “forward problem” is already solved.

1. Choose  $0 \leq \eta \leq 1$  and data point  $\theta \in \{y_1, \dots, y_n\}$ .<sup>6</sup>
2. Perform pre-transformations:
  - (a) Shift data set to get a mean of zero. Shift  $\theta$  by the same amount.
  - (b) Orthogonally transform data to make the covariance matrix diagonal; call this new covariance matrix  $\Sigma$ . Apply this transformation also to  $\theta$ .
  - (c) Transform the data set and  $\theta$  by  $\Sigma^{-\frac{1}{2}}$ .
3. Choose starting values  $r_0, p_0$ :
  - (a) The natural choice for  $r_0$  is  $r_0 = -\frac{\theta}{\|\theta\|}$ .
  - (b) For  $p_0$  one can use 0.25.
4. Iterate in  $i$ :
  - (a) Orthogonally transform  $r_i$  to  $b_1$ ; apply the same transformation to the data set and to  $\theta$ .
  - (b) Applying equations (5.4) and (5.6), compute a value  $\theta_i$  corresponding to  $r_i$  and  $p_i$ .

---

<sup>6</sup>In principle,  $\theta$  is not restricted to be one of the data points but may be any point in  $\mathbb{R}^k$ . Of course, there may be a solution only if  $\eta$  is small enough.

(c) Update  $p$ :

i. Define  $q_i := \log \frac{1-p_i}{p_i}$  so that  $p = 0$  corresponds to  $q = \infty$  and  $p = \frac{1}{2}$  corresponds to  $q = 0$ .

ii. Now we can assume that  $\frac{\|\theta\|}{\|\theta_i\|} \approx \frac{q}{q_i}$ , so we set  $q_{i+1} := q_i \frac{\|\theta\|}{\|\theta_i\|}$ .

iii. Transform back to get  $p_{i+1} = \frac{1}{1+\exp(q_{i+1})}$ .

(d) Update  $r$  by adding the vector  $v_i := \frac{\theta}{\|\theta\|} - \frac{\theta_i}{\|\theta_i\|}$  to it. More precisely, since  $r$  points in the opposite direction of  $\theta$ , we define  $r_{i+1} := -(-r_i + v_i) = r_i - v_i$ .

(e) Repeat from 4a until convergence ( $\|\theta_i - \theta\| < \epsilon$ ).

5. Apply the necessary re-transformations (only needed if value for  $r$  is wanted).

6. If wanted, transform  $p$  according to equation (4.18) (see also the appendix).

There is, in fact, an alternative to this algorithm that is much quicker to implement than the one described: instead of transforming the data and  $r$  in each step, use the complete “forward algorithm” for the *untransformed* data set (above we use only the “core” of the forward algorithm, namely equations (5.4) and (5.6)). However, it turns out that in that case the algorithm is far less efficient than the one described. The (likely) reason for this is that the updating procedure of  $r$  (step 4d above) is a lot more efficient for the standardised data set (as is the case in the described algorithm) than for the original data set.

## B Solving equation (4.18) for $\bar{p}$

Equation (4.18) can be written as

$$\frac{\varphi(\varrho) - \varrho\Phi(-\varrho)}{2(\varphi(\varrho) - \varrho\Phi(-\varrho)) + \varrho} - p = 0, \quad \varrho = \sqrt{F_{\chi^2, k}^{-1}(1 - 2\bar{p})}. \quad (\text{B.1})$$

In this section we present the computations that are necessary to solve (B.1) for  $\bar{p}$  numerically when  $p$  is a given constant. We rely on figure 2 to state that the function of which we have to find the root is increasing and concave. This makes the Newton-Raphson method the natural choice for the root-finding algorithm. First we introduce some convenient notation. Let

$$g(x) := \varphi(x) - x\Phi(-x), \quad G(x) := \frac{g(x)}{2g(x) + x} - p, \quad \varrho(x) := \sqrt{F_{\chi^2, k}^{-1}(x)}, \quad (\text{B.2})$$

where  $x > 0$ . Note that

$$g'(x) = \varphi'(x) - (-x\varphi(-x) + \Phi(-x)) = -x\varphi(x) + x\varphi(x) - \Phi(-x) = -\Phi(-x).$$

We have to solve the equation  $G(\varrho(1 - 2\bar{p})) = 0$  for  $\bar{p}$ . Hence, to apply Newton-Raphson we have to compute the term  $\frac{\partial G}{\partial \bar{p}}G(\varrho(1 - 2\bar{p}))$ . We get

$$\frac{\partial G}{\partial \bar{p}}G(\varrho(1 - 2\bar{p})) = -2G'(\varrho(1 - 2\bar{p}))\varrho'(1 - 2\bar{p}),$$

where

$$G'(x) = \frac{(2g(x) + x)g'(x) - g(x)(2g'(x) + 1)}{(2g(x) + x)^2} = \frac{xg'(x) - g(x)}{(2g(x) + x)^2} = \frac{-\varphi(x)}{(2g(x) + x)^2},$$

$$\varrho'(x) = \frac{(F_{\chi^2, k}^{-1})'(x)}{2\sqrt{F_{\chi^2, k}^{-1}(x)}} = \frac{1}{2F'_{\chi^2, k}(F_{\chi^2, k}^{-1}(x))\sqrt{F_{\chi^2, k}^{-1}(x)}} = \frac{1}{2F'_{\chi^2, k}(\varrho(x)^2)\varrho(x)}.$$

Because

$$F_{\chi^2, k}(x) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \int_0^x t^{\frac{k}{2}-1} e^{-\frac{t}{2}} dt$$

we conclude that

$$F'_{\chi^2, k}(x) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}.$$