Best Approximation under a Convex Paranorm

Guido Moerkotte

TR-2008-007

Abstract

We introduce the q-paranorm, investigate some of its properties. We further give an algorithm which constructs the best linear approximations under the q-paranorm.

1 Introduction

1.1 Scenario

One of the major components of every database management system is the query optimizer. It is responsible for finding the best query evaluation plan possible. In order to find the best plan, the query optimizer generates many alterantive query evaluation plans equivalent to the given query. For each of them it calculates the costs via a cost model and returns the cheapest plan as the solution. Thus it is obvious, that the quality of the resulting plan is truely dependent on the accuracy of the cost estimations. The major input to any cost model are the cardinalities of the intermediate results as produced for example by a selection or join. These have to be estimated. The estimation process is based on data summaries, e.g. statistics, of the original data. In today's DBMSs the de-facto standard data summaries are histograms.

Typically, the buckets within a histogram are approximated by the average of the frequencies contained therein. Obviously, if the data is not uniformely distributed within a bucket, this may result in large estimation errors. Thus it is not surprising, that the use of linear functions to approximate the distribution within a bucket was proposed. In fact, König and Weikum suggested to use linear regression to derive a linear approximation [3]. However, using linear regression has several disadvantages. First, as we will argue below, the L_2 norm used by linear regression is not suitable for the problem on hand. Second, linear regression does not give error bounds.

Therefore, we adopt a new measurement of quality, the Q-paranorm, which allows us to derive useful multiplicative error bounds. This measure will not be a norm in the mathematical sense, but it is a perfect fit for our application scenario of cardinality estimation for query optimization purposes.

The next two subsections discusses several quality metrics and their bounds. The consequence will be that the Q-paranorm is most suited for query optimization.

1.2 Approximations and Measurements of Quality

Consider a given set of pairs (x_i, y_i) for $1 \leq i \leq m$ and an approximation function f. Then there exist several measures for the quality of this approximation. If we denote by $f_i := f(x_i)$, some of the quality measures we find in the literature are:

$$\begin{aligned}
L_1(f) &:= \sum_i |y_i - f_i| \\
L'_1(f) &:= 1/m \sum_i |y_i - f_i| \\
L_2(f) &:= \sqrt{\sum_i (y_i - f_i)^2} \\
L'_2(f) &:= 1/m \sum_i (y_i - f_i)^2 \\
S_1(f) &:= \max_i |y_i - f_i| \\
S_R(f) &:= \max_i |\frac{y_i - f_i}{y_i}| \\
S'_R(f) &:= \max_i |\frac{y_i - f_i}{f_i}| \\
S_Q(f) &:= \max_i \max(\frac{y_i}{f_i}, \frac{f_i}{y_i})
\end{aligned}$$

L_1 is the well-known L1-norm, L_2 is the L2-norm, S_R is the Chebyshev norm or L_∞ norm and S_Q is what we will call the Q-paranorm. The Q-paranorm is not new. It was also used as a quality metrics in the context of distinct sampling [4, 1].

An approximation problem is defined as follows. Given the pairs (x_i, y_i), a quality measure M and a set of functions F, find the function $f \in F$ such that f minimizes M. The set of functions F could be the set of linear functions, polynomials etc.

Clearly, some measures are equivalent in that they yield the same f as a solution (e.g. L_i and L'_i), while others are not.

1.3 Example

Since the authors of the papers are database researchers mainly interested in query optimization, the ultimate purpose of this work (for us) is to approximate buckets of a histogram by some linear combination of functions. Even more, after the general theory, we will restrict ourselves to linear functions. The reason is that this consumes the least space: only two parameters have to be kept per histogram bucket. Another set of functions we are interested in are e^{b+ax} for parameters a and b (see Sec. 4.3).

These approximations are then used by a query optimizer to estimate cardinalitities and calculate costs. With this goal in mind, the kind of approximation and its guarantees play a crucial role. Next, we discuss several alternatives via a simple example.

Consider the three values

$$(1, 20), (2, 10), (3, 60)$$

and assume that they are the only values within a given histogram bucket.

Traditionally, the mean value $\overline{y} = 30$ would be used to approximate estimates within the bucket. The according function is $f_{30}(x) = 30 + 0x$. Therefore, the uniform distribution assumption is used.

König and Weikum use linear regression to gain a linear function $f_{\text{lg}}(x) := \beta + \alpha x$, which is then used to produce estimates within a bucket [3]. Linear regression finds the linear function f_{lg} that minimizes L_2.

Obviously, every measure results in a different function. Let us take the above example and consider as F the set of linear functions $f_{\alpha,\beta}(x) = \alpha + \beta x$. The following table shows the values of x, y and estimates for y for the functions f_{30}, f_{10x}, f_{lg}, f_{S_1}, f_{S_R}, $f_{S'_R}$ and f_{S_Q}. The latter minimize S_1, S_R, S'_R, and S_Q, resp. Additionally, we give the measures L_1, L_2, and S_1 and S_Q for each of these approximations, as well as the α and β.

x	y	f_{30}	f_{lg}	f_{S_1}	f_{S_R}	$f_{S'_R}$	f_{S_Q}
1	20	30	10	5	8	12.5	10
2	10	30	30	25	16	25.0	20
3	60	30	50	45	24	37.5	30
L'_1		20	13	15	18	15	17
L'_2		467	200	225	492	262.5	367
S_1		30	20	15	36	22.5	30
S_R		2	2	1.5	0.6	1.5	1
S'_R		1	1	3	1.5	0.6	1
S_Q		3	3	4	2.5	2.5	2
β		30	-10	-15	0	0	0
α		0	20	20	8	12.5	10

The question arises, which is the best norm for cardinality estimation purposes. A very nice property of a quality measure is the existance of lower and upper bounds. Thereby we mean by a bound the following. Assume we have an estimate f_i for some real value y_i. Knowing only f_i, we would like to infer an interval to which y_i surely belongs. This interval can of course only be derived if we know something more than only the f_i. This something more is what we call the error bounds. For the Chebyshev norm, it could be the maximal deviation d of all estimates within a bucket. Then we know that $f_i - d \leq y_i \leq f_i + d$. We give these kinds of lower and upper bounds for the above quality metrics.

Using this information then allows the plan generator to perform a sensitivity analysis of the generated plans. Thus, error bounds are the subject studied in the next section.

1.4 Error Bounds

Those measures, which somehow average (for example L_1, L_2) do not allow us to derive error bounds. Those measures, which use max (for example S_1, S_R, S_Q), do allow us to derive error bounds. More specifically, for a given estimate it is possible to derive an interval, which for sure contains the true value. In the following we derive these error bounds for S_1, S_R, S'_R, and S_Q.

Observation 1 **(S_1)** *Assume f is given and $c := S_1(f)$ is known. Then*

$$f_i - c \leq y_i \leq f_i + c$$

$-c$ and $+c$ are called the determinators *for the lower and upper error bound.*

Proof: From $c = \max_i |y_i - f_i|$ is follows that $|y_i - f_i| \leq c$. For $y_i \geq f_i$ we thus have $y_i - f_i \leq c$ and, hence, $y_i \leq f_i + c$ and, thus, $f_i - c \leq y_i \leq f_i + c$. For $y_i \leq f_i$ we thus have $f_i - y_i \leq c$ and, hence, $f_i - c \leq y_i$ and, thus, $f_i - c \leq y_i \leq f_i + c$. □

Observation 2 **(S_R)** *Assume f is given and $c := S_R(f)$ is known. Then*

$$\frac{1}{c+1} f_i \leq y_i \leq \frac{1}{1-c} f_i$$

$1/(1+c)$ and $1/(1-c)$ are called the determinators *for the lower and upper error bound.*

Proof: For $y_i \geq f_i$ it follows that

$$\begin{array}{lll} & (y_i - f_i)/y_i \leq c & \\ \Longrightarrow & y_i - f_i \leq c y_i & \\ \Longrightarrow & (1-c) y_i \leq f_i & ll \\ \Longrightarrow & y_i \leq \frac{1}{1-c} f_i & \\ \Longrightarrow & \frac{1}{1+c} f_i \leq y_i \leq \frac{1}{1-c} f_i & \end{array}$$

For $y_i \leq f_i$ it follows that

$$\begin{array}{lll} & (f_i - y_i)/y_i \leq c & \\ \Longrightarrow & f_i \leq (1+c) y_i & ll \\ \Longrightarrow & \frac{1}{c+1} f_i \leq y_i & \\ \Longrightarrow & \frac{1}{1+c} f_i \leq y_i \leq \frac{1}{1-c} f_i & \end{array}$$

□

Observation 3 **(S'_R)** *Assume f is given and $c := S'_R(f)$ is known. Then*

$$(1-c) f_i \leq y_i \leq (1+c) f_i$$

$1-c$ and $1+c$ are called the determinators *for the lower and upper error bound.*

Proof: For $y_i \geq f_i$ it follows from $|y_i - f_i| \leq cf_i$ and $0 \leq c$ that

$$\begin{aligned} & y_i - f_i \leq cf_i \\ \Longrightarrow \quad & (1-c)f_i \leq f_i \leq y_i \leq (1+c)f_i \end{aligned}$$

For $y_i \leq f_i$ it follows from $|y_i - f_i| \leq cf_i$ and $0 \leq c$ that

$$\begin{aligned} & f_i - y_i \leq cf_i \\ \Longrightarrow \quad & (1-c)f_i \leq y_i \leq f_i \leq (1+c)f_i \end{aligned}$$

□

Observation 4 **(S_Q)** *Assume f is given and $c := S_Q(f)$ is known. Then*

$$1/cf_i \leq y_i \leq cf_i$$

$1/c$ and c are called the determinators *for the lower and upper error bound.*

Proof: For $y_i \geq f_i$ it follows that

$$\begin{aligned} & c \geq y_i/f_i \geq 1 \geq f_i/y_i \\ \Longrightarrow \quad & 1/cf_i \leq f_i \leq y_i \leq cf_i \end{aligned}$$

For $y_i \leq f_i$ it follows that

$$\begin{aligned} & c \geq f_i/y_i \geq 1 \geq y_i/f_i \\ \Longrightarrow \quad & 1/cf_i \leq y_i \leq f_i \leq cf_i \end{aligned}$$

□

For our introductory example, the determinators for the lower and upper error bound are given in the following table:

S_1		S_R		S'_R		S_Q	
low	high	low	high	low	high	low	high
-15	15	0.625	2.5	0.4	1.6	0.5	2.0

Observe that S_1 and S_Q are the only symmetric error bounds. As cardinality estimates are multiplied with other numbers to derive costs of an operator, a symmetric error bound is nice to have, or, to quote Charikar et al.[4]:

> "We do not favor the use of relative error as it is fairly misleading in comparing an overestimate with an underestimate."

Further, minimizing absolute errors as under S_1 is less useful under this scenario than optimizing relative errors. If the bucket has frequencies in the interval of $[10, 100]$, an error bound of 10 corresponds to a relative error between 10% and 100%.

Or remember our introductory example. For the optimal approximation under the Chebyshev norm this then gives the maximal absolute deviation of -15 and 15 for all of our three points. Thus, from $f_1 = 5$ we can derive that $y_1 \in [-10, 20]$, from $f_2 = 25$ it follows that $y_2 \in [10, 40]$ and for $f_3 = 45$ we infer that $y_3 \in [30, 60]$. Obviously, from a query optimizers point of view the deviation for y_1 is much worse than that for y_3. In many cost model for operators, the input cardinality determines the costs linearly. That is, if the input is twice as big, the cost for the selection are twice as high. If one of the input relations of a join is half as big, the costs for the join or its memory consumption typically half. What really counts for cardinality estimation is the relative deviation as measured by the Q-paranorm. Another strong argument for the Q-paranorm is error propagation. As Ioannidis and Christodoulakis pointed out, errors propagate multiplicatively through joins [2]. Assume we want to join three relations R_1, R_2, and R_3 and that the cardinality estimates of R_i are each a factor of 2 off. Then, the cardinality estimation of $R_1 \bowtie R_2 \bowtie R_3$ will be a factor of 8 off. Hence, minimizing the multiplicative error also minimizes the propagated error.

1.5 Contribution

Our contributions are theoretical results for the Q-paranorm and an algorithm that is capable to construct optimal approximations under S_Q.

2 The Convex Paranorm $||\cdot||_Q$ and its Basic Properties

2.1 Q-paranorm in R

Definition 1 (Q-paranorm in R) *Define for $x \in R$*

$$||x||_Q = \begin{cases} \infty & \text{if } x \leq 0 \\ \max(x, 1/x) & \text{else} \end{cases}$$

$||\cdot||_Q$ *is called* Q-paranorm.

Definition 2 (norm) *Let S be a linear space. Then a function $||x|| : S \to R$ is called a* norm *if and only if it has the following three properties:*

1. $||x|| > 0$ *unless* $x = 0$
2. $||\lambda x|| = |\lambda| \; ||x||$
3. $||x + y|| \leq ||x|| + ||y||$

Observation 5 $||\cdot||_Q$ *is not a norm since 1) and 2) do not hold. However, 3 does hold.*

Proof:
ad 1) obvious.
ad 2) consider $\lambda = 1/2$, $x = 1/4$: $8 = 1/(1/2 * 1/4) = ||\lambda x||_Q > \lambda ||x||_Q = 1/2 * 4 = 2$.
ad 3): We consider the following cases:

- $x = 0 \vee y = 0$ $\surd$
- $x \geq 1$, $y \geq 1$ ($\Longrightarrow x + y \geq 1$)
  $x + y \leq x + y$ $\surd$
- $x < 1$, $y < 1$
  - $x + y < 1$
    $\frac{1}{x+y} \leq \frac{1}{x} + \frac{1}{y}$ $\surd$
  - $x + y \geq 1$
    $x + y \leq \frac{1}{x} + \frac{1}{y}$ $\surd$
- $x < 1$, $y \geq 1$ ($\Longrightarrow x + y \geq 1$)
  $x + y \leq \frac{1}{x} + y$ $\surd$
- $x \geq 1$, $y < 1$: by symmetry

□

Since the second condition does not hold, let us consider the different cases explicitly:

case 1 $x < 1$, $y < 1$: $||xy||_Q = ||x||_Q ||y||_Q = 1/x||y||_Q = 1/y||x||_Q$

case 2 $x \geq 1$, $y \geq 1$: $||xy||_Q = ||x||_Q ||y||_Q = xy = x||y||_Q = y||x||_Q$

case 3a $x \geq 1$, $y < 1$, $xy \leq 1$: $||xy||_Q = 1/x||y||_Q = 1/y||x||_Q$

case 3b $x \geq 1$, $y < 1$, $xy > 1$: $||xy||_Q = xy = y||x||_Q$

Definition 3 (paranorm) *Let S be a linear space. Then a function $||x|| : S \to R$ is called a* paranorm *if and only if the following two properties hold:*

1. $||x|| \geq 0$
2. $||x + y|| \leq ||x|| + ||y||$

The following lemma is an immediate consequence of the definition of paranorm and the above considerations:

Lemma 1 $|| \cdot ||_Q$ *is a paranorm.*

As the convexity of a function plays an important role in approximation theory, we repeat its definition.

Definition 4 *A function f is* convex *if and only if for all x, y and $0 \leq \lambda \leq 1$:*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y) \tag{1}$$

Observation 6 *For $0 < y < x < z$ and $0 < z < x < y$ we have $||x/z||_Q < ||y/z||_Q$.*

Proof: $0 < y < x < z$: First note that this implies $x/z < 1$ and $y/z < 1$. Then

$$\begin{array}{rrcl} \Longrightarrow & y/z & < & x/z \\ \Longrightarrow & z/x & < & z/y \\ \Longrightarrow & ||x/z||_Q & < & ||y/z||_Q \end{array}$$

$0 < z < x < y$: First note that this implies $x/z > 1$ and $y/z > 1$. Then

$$\begin{array}{rrcl} \Longrightarrow & x/z & < & y/z \\ \Longrightarrow & ||x/z||_Q & < & ||y/z||_Q \end{array}$$

□

Lemma 2 $|| \cdot ||_Q$ *on R is convex.*

proof case 1: $\lambda x + (1-\lambda)y \geq 1$ ($\Longrightarrow x \geq 1 \vee y \geq 1$)
case 1.1: $x \geq 1 \wedge y \geq 1$ ($\Longrightarrow ||x||_Q = x$, $||y||_Q = y$)
$\max(\cdot) = \lambda x + (1-\lambda)y \leq \lambda x + (1-\lambda)y \surd$
case 1.2: $x \geq 1 \wedge y < 1$ ($\Longrightarrow ||x||_Q = x$, $||y||_Q = 1/y$)
$\max(\cdot) = \lambda x(1-\lambda)y \leq \lambda x + (1-\lambda)1/y \surd$
case 1.3: $x < 1 \wedge y \geq 1$: by symmetry.
case 2: $\lambda x + (1-\lambda)y < 1$ ($\Longrightarrow x < 1 \vee y < 1$)
First note that $x + 1/x$ has minimum 2 at $x = 1$.
case 2.1: $x < 1$, $y < 1$ ($\Longrightarrow ||x||_Q = 1/x$, $||y||_Q = 1/y$)

$$\begin{aligned} 1 &\leq 1 - 2\lambda + 2\lambda^2 + \lambda(1-\lambda)(x/y + y/x) \\ &\leq \lambda^2 + \frac{(1-\lambda)x}{y} + \frac{\lambda(1-\lambda)y}{x} + (1-\lambda)^2 \\ &\leq (\lambda x + (1-\lambda)y)(\lambda\frac{1}{x} + (1-\lambda)\frac{1}{y}) \end{aligned}$$

case 2.2: $x < 1$, $y \geq 1$ ($\Longrightarrow ||x||_Q = 1/x$, $||y||_Q = y$)
From $1/x > 1$, $y \geq 1$ is follows:

$$\begin{aligned} 1 &\leq \lambda^2 + y[\lambda(1-\lambda)(1/x + x) + (1-\lambda)^2 y] \\ &\leq \lambda^2 + y[\lambda(1-\lambda)1/x + \lambda(1-\lambda)x + (1-\lambda)^2 y] \\ &\leq \lambda^2 + \lambda(1-\lambda)y/x + \lambda(1-\lambda)xy + (1-\lambda)^2 y^2 \\ &\leq \lambda x(\lambda 1/x + (1-\lambda)y) + (1-\lambda)y(\lambda 1/x + (1-\lambda)y) \end{aligned}$$

case 2.3: $x \geq 1$, $y < 1$: by symmetry.

□

2.2 Q-paranorm in R^n

Definition 5 (Q-paranorm in R^n) *For $x \in R^n$, $x_i \neq 0$, define $||x||_Q = \max_i ||x||_Q$*

Lemma 3 *$||\cdot||_Q$ on R^n is a paranorm.*

Lemma 4 *$||\cdot||_Q$ on R^n is convex.*

3 The Problem and its Solution

This section is organized as follows. After presenting some preliminaries, we formally define the problem. Next, we show the existence of a solution and characterize it. Then, we show the uniqueness of the solution for problems fulfilling a certain property. Finally, we show how to derive the solution constructively.

3.1 Preliminaries

We need some preliminaries (see Chapter 1 of [5]). But first note that in this paper all (non-transposed) vectors are column vectors.

Definition 6 (convex hull) *Let $D \subset R^n$. Then we define the* convex hull *of D as*

$$conv(D) = \{d | d = \sum_i \lambda_i d_i, d_i \in D, \sum_i \lambda_i = 1, \lambda_i \geq 0\}$$

where the only restriction for the sums is that they have to be finite.

Theorem 1 (Caratheodory's theorem, cmp. Theorem 1.4 of [5]) *Let $D \subset R^n$. Then any $h \in conv(D)$ can be expressed as a linear combination of $(n+1)$ or fewer points.*

Theorem 2 (Theorem 1.5 of [5]) *Let D be a closed, convex subset of R^n. Then D does not contain the origin if and only if there exists $z \in R^n$ such that*

$$d^T z > 0$$

for all $d \in D$.

3.2 Problem

Let a and b be two vectors in R^n with $b_i \neq 0$. Then, we define $a/b = \frac{a}{b} = (a_1/b_1, \ldots, a_n/b_n)^T$.

Problem 1 *Let $b \in R^m$ be a vector with $b_i > 0$ and A an $m \times n$ matrix. Denote by α_i the vector formed from the i-th row of A. The problem is defined as follows:*

$$\textit{Find } a \in R^n \textit{ to minimize } ||Aa/b||_Q$$

under the constraint that $\alpha_i^T a > 0$ *for all* $1 \leq i \leq m$. □

For the rest of this section, we assume that n, m, b, and A (α_i) are given such that the conditions of Problem 1 are satisfied. Additionally, we define a *quotient vector* $q(a) = (\alpha_1^T a/b_1, \ldots, \alpha_m^T a/b_m)^T$ and the *residual vector* $r(a) = b - Aa$. The components of $q(a)$ $(r(a))$ are denoted by $q_i(a)$ $(r_i(a))$. If a is understood from context, we may omit it.

3.3 Existence of Solution

The problem states that we have to find an approximation for b in the subspace M defined by Aa for $a \in R^n$ with $\alpha_i^T a > 0$. A necessary condition for the existence of a solution is that M is compact, which is obviously not the case here. However, we can construct a compact subspace. Let c be an upper bound for $min_a||Aa/b||_Q$. Then, it suffices to find an approximation in the compact subspace $\{Aa|\ ||Aa/b||_Q \leq c\}$.

For norms the existence of a solution on compact spaces is guaranteed. As $||\cdot||_Q$ is not a norm, we need to prove the existence of a solution to Problem 1.

The first theorem is the analog of Theorem 1.1 of [5] but instead of requiring a norm, we use our paranorm $||\cdot||_Q$.

Theorem 3 *Let M be a compact subset of R^n and for each $x \in M$ and each $1 \leq i \leq n$ $x_i > 0$. Then, for each point $g \in R^n$, $g_i > 0$, exists a point $a \in M$ such that $||a/g||_Q = min_{a \in M}||a/g||_Q$.*

Proof: Let $\delta = \inf\{||x/g||_Q | x \in M\}$. By the definition of infimum, there exists a sequence (x_i), $x_i \in M$, such that

$$||x_i/g||_Q \rightarrow \delta$$

for $i \rightarrow \infty$. Since M is compact, there exists a subsequence of (x_i) converging to $x^* \in M$. Now

$$\begin{aligned} ||x^*/g||_Q &= ||\frac{x_i - (x_i - x^*)}{g}||_Q \\ &= ||\frac{x_i}{g} - \frac{x_i - x^*}{g}||_Q \\ &\leq ||x_i/g||_Q + ||x_i - x^*/g||_Q \end{aligned}$$

and, for $i \to \infty$ it follows that $||x^*/g||_Q \leq \delta$. Since $x^* \in M$, we also have $||x^*/g||_Q \geq \delta$ and thus $||x^*/g||_Q = \delta$, which shows the claim. □

3.4 Convexity-based Approach to Uniqueness of the Solution

This section mainly shows that a simple often persued approach to uniqueness does not work for our paranorm. It is intended for readers who wonder why things are as complicated as they appear.

Definition 7 (closed q-sphere) *For $a \in R^n$, the set*

$$\{x| \ ||x/b||_Q \leq r\}$$

is called a close q-sphere *with radius r and center b.*

Let $b = 10$ and $r = 2$, then the q-sphere with radius r and center b is

$$[-20, -5] \cup [5, 20] \cup \{0\}$$

Definition 8 (q-convex set) *A subset M of R^n is convex if $x, y \in M$, and $(x > 0 \wedge y > 0) \vee (x < 0 \wedge y < 0)$ implies that $\lambda x + (1-\lambda)y \in M$ for all $0 \leq \lambda \leq 1$.*

Lemma 5 *Closed q-spheres are q-convex.*

Proof: Denote by $S \subset R^n$ a sphere with radius r and center $b \in R^n$. Let $x, y \in S$ and λ with $0 \leq \lambda \leq 1$. Then

$$\begin{aligned} ||\frac{\lambda x_1 + (1-\lambda)x_2}{b}||_Q &= ||\frac{\lambda x_1}{b} + \frac{(1-\lambda)x_2}{b}||_Q \\ &\leq \lambda ||x_1/b||_Q + (1-\lambda)||x_2/b||_Q \\ &\leq r \end{aligned}$$

where we used the convexity of $|| \cdot ||_Q$. □

Definition 9 (strictly q-convex paranormed linear space) *Let S be a linear space. A convex function $f : S \to R$ is* strictly q-convex *if and only* *if for all points* $x, y \in S$*,* $x \neq y$*,* $(x > 0 \wedge y > 0) \vee (x < 0 \wedge y < 0)$*, on the boarder of a sphere with center* b *and radius* y *the following holds:*

$$\begin{aligned} f(\lambda x + (1-\lambda)y - b) &< r \quad \textit{alternative} \\ f((\lambda x + (1-\lambda)y)/b) &< r \end{aligned}$$

for all $0 < \lambda < 1$*. If in a paranormed linear space the paranorm is strictly q-convex, then we have a strictly q-convex paranormed linear space.*

Theorem 4 *In a strictly q-convex paranormed linear space* S *a finite dimensional subspace* M *contains a unique best positive and a unique best negative approximation to any point* $g \in S$*,* $g_i \geq 0$ *or* $g_i \leq 0$*.*

Lemma 6 R *with* $||\cdot||_Q$ *is a strictly q-convex paranormed linear space.*

Proof: Let $x, y \in R$, $x \neq y$ be two points on a sphere with center b and radius a. Then $||x/b||_Q = ||y/b||_Q$ $x \neq y$ implies that □

Observation 7 R^n *with* $||\cdot||_Q$ *is* not *a strictly convex paranormed linear space for* $n > 1$*.*

From this, it follows that we cannot use the approach of Theorem 4 to show the uniqueness of a solution to Problem 1.

3.5 Characterization of the Solution

Let us denote by $\bar{I}(a) \subseteq \{1, \ldots, m\}$ the set of indices i corresponding to those components of $q(a)$ with $||q_i(a)||_Q = ||q(a)||_Q$. Obviously, $\bar{I}(a)$ is not empty. Define

$$\theta_i(a) = \text{sign}(r_i(a))$$

Let $\theta(a)$ be the vector formed from the $\theta_i(a)$ for $1, \ldots, m$. If a is clear from context, we omit it.

For the example of section 1.3, we have

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}$$

and $b = (20, 10, 60)^T$. The function f_{S_Q} results from solution $a = (0, 10)^T$. We get

$$r(a) = \begin{pmatrix} 20 \\ 10 \\ 60 \end{pmatrix} - \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 0 \\ 10 \end{pmatrix} = \begin{pmatrix} 10 \\ -10 \\ 30 \end{pmatrix}$$

and

$$q(a) = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 0 \\ 10 \end{pmatrix} / \begin{pmatrix} 20 \\ 10 \\ 60 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 2 \\ 0.5 \end{pmatrix}$$

Thus, $\theta(a) = (+1, -1, +1)^T$ and $\bar{I}(a) = \{1, 2, 3\}$.

Let us do some very simple calculations to get used to the notation. For positive numbers and a given vector a, we have

$$\begin{aligned} ||q_i||_Q &= ||\frac{\alpha_i^T a}{b_i}||_Q \\ &= (\frac{\alpha_i^T a}{b_i})^{-\theta_i} \\ &= q_i^{-\theta_i} \end{aligned}$$

and thus $||q_i||_Q^{-\theta_i} = q_i$. Using this, we can easily derive that

$$\begin{aligned} q_i &= ||q_i||_Q^{-\theta_i} \\ b_i &= \frac{\alpha_i^T a}{q_i} = ||q_i||_Q^{\theta_i} \alpha_i^T a \\ b_i &= ||q||_Q^{\theta_i} \alpha_i^T a \quad \text{for } i \in \bar{I} \\ \alpha_i^T a &= ||q_i||_Q^{-\theta_i}\ b_i \\ r_i(a) &= b - \alpha_i^T a = b_i - ||q_i(a)||_Q^{-\theta_i(a)}\ b_i = (1 - ||q_i(a)||_Q^{-\theta_i(a)})\ b_i \\ r_i(a) &= (1 - ||q(a)||_Q^{-\theta_i(a)})\ b_i \quad \text{for } i \in \bar{I} \end{aligned}$$

$|r_i|$ and $||q_i||_Q$ are not directly related for two different vectors. But some relationships hold as the following two observations show:

Observation 8 *Let a and c be two vectors in R^n with*

$$||q_i(a)||_Q \leq ||q_i(c)||_Q.$$

Then

a *If* $\theta_i(a) \leq 0$*,* $\theta_i(c) \leq 0$ *then* $|r_i(a)| \leq |r_i(c)|$

b *If* $\theta_i(a) \geq 0$*,* $\theta_i(c) \geq 0$ *then* $|r_i(a)| \leq |r_i(c)|$

c *If* $\theta_i(a) > 0$*,* $\theta_i(c) \leq 0$ *then* $|r_i(a)| \leq |r_i(c)|$

d *If* $\theta_i(a) < 0$*,* $\theta_i(c) \geq 0$ *then* $|r_i(a)| \geq |r_i(c)|$

where equality only holds if $||q_i(a)||_Q = ||q_i(c)||_Q$*.*

Proof: Assume $||q_i(a)||_Q \leq ||q_i(c)||_Q$ (*) and consider the following 4 cases:

a $\frac{\alpha_i^T a}{b_i} \geq 1$, $\frac{\alpha_i^T c}{b_i} \geq 1$
$\Longrightarrow \alpha_i^T a \geq b_i$, $\alpha_i^T c \geq b_i$ (**)

$$\begin{array}{lrcll}
(*) \Longrightarrow & \frac{\alpha_i^T a}{b_i} & < & \frac{\alpha_i^T c}{b_i} & \\
\Longrightarrow & \alpha_i^T a & < & \alpha_i^T c & \\
(**) \Longrightarrow & b_i - \alpha_i^T a & > & b_i - \alpha_i^T c & \text{(both sides negative!)} \\
\Longrightarrow & |b_i - \alpha_i^T a| & < & |b_i - \alpha_i^T c| & \\
\Longrightarrow & |r_i(a)| & < & |r_i(c)| &
\end{array}$$

b $\frac{\alpha_i^T a}{b_i} \leq 1$, $\frac{\alpha_i^T c}{b_i} \leq 1$
$\Longrightarrow \alpha_i^T a \leq b_i$, $\alpha_i^T c \leq b_i$ (**)

$$\begin{array}{lrcll}
(*) \Longrightarrow & \frac{b_i}{\alpha_i^T a} & \leq & \frac{b_i}{\alpha_i^T c} & \\
\Longrightarrow & \alpha_i^T c & \leq & \alpha_i^T a & \\
\Longrightarrow & b_i + \alpha_i^T c & \leq & b_i + \alpha_i^T a & \\
\Longrightarrow & b_i - \alpha_i^T a & \leq & b_i - \alpha_i^T c & \text{(both sides positive!)} \\
\Longrightarrow & |r_i(a)| & \leq & |r_i(c)| &
\end{array}$$

c $\frac{\alpha_i^T a}{b_i} < 1$, $\frac{\alpha_i^T c}{b_i} \geq 1$: In this case, we have $\theta_i(a) = +1$ and $\theta_i(c) = -1$ or $\theta_i(c) = 0$. The latter implies also $|r_i(c)| = 0$ and thus $|r_i(a)| = 0$. Hence, consider $\theta_i(c) = -1$.

$$(*) \Longrightarrow \frac{\alpha_i^T a}{b_i} < \frac{b_i}{\alpha_i^T a} < \frac{\alpha_i^T c}{b_i}$$

For $x \geq 1$, $y \geq 1$, $x < y$ we have $\frac{1}{x} > \frac{1}{y}$ and thus $2 < y + \frac{1}{y} < y + \frac{1}{x}$ from which follows that

$$\begin{array}{lrcll}
 & 2 & < & y + \frac{1}{x} & \\
\Longrightarrow & 1 - \frac{1}{x} & < & y - 1 & \\
\Longrightarrow & |1 - \frac{1}{x}| & < & |1 - y| & (***)
\end{array}$$

From this we can infer that

$$\begin{array}{rcl} |r_i(a)| & = & |(1 - ||q_i(a)||_Q^{-\theta_i(a)}|\ |b_i| \\ & = & |(1 - ||q_i(a)||_Q^{-1}|\ |b_i| \\ & < & |(1 - ||q_i(c)||_Q^{+1}|\ |b_i| \\ & = & |(1 - ||q_i(c)||_Q^{-\theta_i(c)}|\ |b_i| \\ & = & |r_i(c)| \end{array}$$

d $\frac{\alpha_i^T a}{b_i} > 1$, $\frac{\alpha_i^T c}{b_i} \leq 1$: In this case, we have $\theta_i(a) = -1$ and $\theta_i(c) = +1$ or $\theta_i(c) = 0$. The latter case implies $|r_i(c)| = 0$ and thus $|r_i(a)| = 0$. Hence, consider $\theta_i(c) = +1$. Using (***) we can infer that

$$\begin{array}{rcl} |r_i(a)| & = & |(1 - ||q_i(a)||_Q^{-\theta_i(a)}|\ |b_i| \\ & = & |(1 - ||q_i(a)||_Q^{+1}|\ |b_i| \\ & > & |(1 - ||q_i(c)||_Q^{-1}|\ |b_i| \\ & = & |(1 - ||q_i(c)||_Q^{-\theta_i(c)}|\ |b_i| \\ & = & |r_i(c)| \end{array}$$

□

Let us give an example for the last case. Define

$$b = \begin{pmatrix} 5 \\ 10 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad a = \begin{pmatrix} 0 \\ 10 \end{pmatrix}, \quad c = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

Then

$$Aa = \begin{pmatrix} 10 \\ 20 \end{pmatrix}, \quad Ac = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad r(a) = b - Aa = \begin{pmatrix} -5 \\ -10 \end{pmatrix}, \quad r(c) = b - Ac = \begin{pmatrix} 3 \\ 8 \end{pmatrix}$$

and

$$q(a) = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad q(c) = \begin{pmatrix} 2/5 \\ 1/5 \end{pmatrix}$$

Thus

$$\begin{array}{rcl} ||q_2(a)||_Q = 2 & < & 5 = ||q_2(c)||_Q \\ |r_2(a)| = 10 & > & 8 = |r_2(c)| \end{array}$$

Let us now look at the opposite direction.

Observation 9 *Let a and c be two vectors in R^n with*

$$|r_i(a)| \leq |r_i(c)|$$

Then

a *If $\theta_i(a) \geq 0$, $\theta_i(c) \geq 0$ then $||\frac{\alpha_i^T a}{b_i}||_Q \leq ||\frac{\alpha_i^T c}{b_i}||_Q$*

b *If $\theta_i(a) \leq 0$, $\theta_i(c) \leq 0$ then $||\frac{\alpha_i^T a}{b_i}||_Q \leq ||\frac{\alpha_i^T c}{b_i}||_Q$*

c *If $\theta_i(a) > 0$, $\theta_i(c) \leq 0$ then nothing can be said.*

d *If $\theta_i(a) < 0$, $\theta_i(c) \geq 0$ then $||\frac{\alpha_i^T c}{b_i}||_Q \leq ||\frac{\alpha_i^T a}{b_i}||_Q$*

where equality only holds if $|r_i(a)| = |r_i(c)|$

Proof: Assume $|r_i(a)| \leq |r_i(c)|$ and consider the following 4 cases:

a $r_i(a) \geq 0$, $r_i(c) \geq 0$ ($\Longrightarrow b_i \geq \alpha_i^T a$, $b_i \geq \alpha_i^T c$)

$$\begin{array}{rrcr} & |r_i(a)| & \leq & |r_i(c)| \\ \Longrightarrow & r_i(a) & \leq & r_i(c) \\ \Longrightarrow & b_i - \alpha_i^T a & \leq & b_i - \alpha_i^T c \\ \Longrightarrow & -\alpha_i^T a & \leq & -\alpha_i^T c \\ \Longrightarrow & \alpha_i^T c & \leq & \alpha_i^T a \\ \Longrightarrow & \frac{\alpha_i^T c}{b_i} & \leq & \frac{\alpha_i^T a}{b_i} \\ \Longrightarrow & \frac{b_i}{\alpha_i^T a} & \leq & \frac{b_i}{\alpha_i^T b} \\ \Longrightarrow & ||\frac{\alpha_i^T a}{b_i}||_Q & \leq & ||\frac{\alpha_i^T c}{b_i}||_Q \end{array}$$

b $r_i(a) \leq 0$, $r_i(c) \leq 0$ ($\Longrightarrow b_i \leq \alpha_i^T a$, $b_i \leq \alpha_i^T c$)

$$\begin{array}{rrcr} & |r_i(a)| & \leq & |r_i(c)| \\ \Longrightarrow & -r_i(a) & \leq & -r_i(c) \\ \Longrightarrow & -b_i + \alpha_i^T a & \leq & -b_i + \alpha_i^T c \\ \Longrightarrow & \alpha_i^T a & \leq & \alpha_i^T c \\ \Longrightarrow & \frac{\alpha_i^T a}{b_i} & \leq & \frac{\alpha_i^T c}{b_i} \\ \Longrightarrow & ||\frac{\alpha_i^T a}{b_i}||_Q & \leq & ||\frac{\alpha_i^T c}{b_i}||_Q \end{array}$$

c $r_i(a) > 0$, $r_i(c) \leq 0$ ($\Longrightarrow b_i \geq \alpha_i^T a$, $b_i < \alpha_i^T c$)

In this case nothing can be said (see example below).

d $r_i(a) < 0$, $r_i(c) \geq 0$ ($\Longrightarrow b_i \leq \alpha_i^T a$, $b_i \geq \alpha_i^T c$)

$$\begin{array}{rrcr} & b_i - |r_i(a)| & \geq & b_i - |r_i(c)| \\ \Longrightarrow & b_i - |r_i(c)| & \leq & b_i - |r_i(a)| \\ \Longrightarrow & b_i - (b_i - \alpha_i^T c) & \leq & b_i - (-b_i + (\alpha_i^T a)) \\ \Longrightarrow & \alpha_i^T c & \leq & 2b_i + \alpha_i^T a \\ \Longrightarrow & ||\frac{\alpha_i^T c}{b_i}||_Q & \leq & ||\frac{\alpha_i^T a}{b_i}||_Q \end{array}$$

The last step follows from the fact that for $x \geq 1$ we have that $2 \leq x + 1/x$. This immediately follows from the fact that the function $f(x) = x + 1/x$ is strongly monotonically increasing and $f(1) = 2$.

Let us give an example for case c: Let $b_i = 10$. From $\alpha_i^T a = 9$ and $\alpha_i^T c = 15$, it follows that $r_i(a) = 1$ and $r_i(c) = -5$ and

$$||\frac{\alpha_i^T a}{b_i}||_Q = \frac{10}{9} < \frac{15}{10} = ||\frac{\alpha_i^T c}{b_i}||_Q$$

On the other hand, if $\alpha_i^T a = 6$ and and $\alpha_i^T c = 15$, it follows that $r_i(a) = 4$ and $r_i(c) = -5$ and

$$||\frac{\alpha_i^T a}{b_i}||_Q = \frac{10}{6} > \frac{15}{10} = ||\frac{\alpha_i^T c}{b_i}||_Q$$

Next is the central theorem. It is very important in that it will be used in almost every of the following proofs.

Theorem 5 *The vector $a \in R^n$ solves Problem 1 if and only if there exists $I \subseteq \bar{I}(a)$, $|I| \leq n + 1$ and $\lambda \in R^m$, $\lambda \neq \vec{0}$ such that the following holds:*

1. *$\lambda_i = 0$ for all $i \notin I$*

2. *$A^T \lambda = \vec{0}$*

3. *$\lambda_i \theta_i \geq 0$ for all $i \in I$*

Note that we can always make I smaller such that (3) can be replaced by $\lambda_i \theta_i > 0$ for all $i \in I$. Further note that (2) is equivalent to $\sum_{i \in I} \lambda_i \alpha_i = \vec{0}$. (Remember that α_i is the i-th row of A.)

Proof: "$\Longleftarrow$": Suppose the conditions are true but a is not a solution. Then there exists a vector $c \in R^n$ such that

$$||q(a + c)||_Q < ||q(a)||_Q$$

In particular, this implies for all $i \in \bar{I}$:

$$||\frac{\alpha_i^T a}{b_i} + \frac{\alpha_i^T c}{b_i}||_Q < ||\frac{\alpha_i^T a}{b_i}||_Q$$

If $\alpha_i^T a/b_i > 1$ then it follows that $\alpha_i^T c < 0$ and thus $\theta_i \alpha_i^T c > 0$.

If $\alpha_i^T a/b_i < 1$ and

$$\frac{\alpha_i^T a}{b_i} + \frac{\alpha_i^T c}{b_i} < 1$$

it follows that

$$\frac{b_i}{\alpha_i^T a} < \frac{b_i}{\alpha_i^T a} + \frac{b_i}{\alpha_i^T c}$$

and, thus, $\alpha_i^T c > 0$ [since $\alpha_i^T a > 0$] and, hence, $\theta_i \alpha_i^T c > 0$.

If $\alpha_i^T a/b_i > 1$ and

$$\frac{\alpha_i^T a}{b_i} + \frac{\alpha_i^T c}{b_i} > 1$$

it follows that

$$\frac{b_i}{\alpha_i^T a} < \frac{\alpha_i^T a}{b_i} + \frac{\alpha_i^T c}{b_i}$$

and, thus, $\alpha_i^T c > 0$ [since $b_i/\alpha_i a > 1$ and $\alpha_i a/b_i < 1$] and, hence, $\theta_i \alpha_i^T c > 0$.

Summarizing, we always have

$$\theta_i \alpha_i^T c > 0.$$

The conditions of the theorem give us

$$\begin{array}{lrcl} \sum_{i \in I} \lambda_i \alpha_i^T & & = & \vec{0} \\ \Longrightarrow & \sum_{i \in I} \lambda_i \theta_i \theta_i \alpha_i^T c & = & \vec{0} \\ \Longrightarrow & \sum_{i \in I} (\lambda_i \theta_i)(\theta_i \alpha_i^T c) & = & \vec{0} \end{array}$$

Since $\lambda_i \theta_i \geq 0$ and $\theta_i \alpha_i^T c > 0$, we have constructed a contradiction to the fact that λ is non-trivial.

"$\Longrightarrow$": Let a be a solution to Problem 1. Let D be the convex hull of $\{\theta_i \alpha_i | i \in \bar{I}(a)\}$. Assume $\vec{0} \notin D$. Then, according to Theorem 2, there exists a $c \in R^n$ such that

$$\theta_i \alpha_i^T c > 0$$

for all $i \in \bar{I}(a)$. $\forall i \in \bar{I}(a)$ and $\forall \gamma$ we have

$$||q_i(a + \gamma c)||_Q = ||\frac{\alpha_i a}{b_i} + \frac{\gamma \alpha_i c}{b_i}||_Q$$

Case 1: $i \notin \bar{I}(a)$: Define

$$\Delta = ||q_a||_Q - \max_{i \notin \bar{I}(a)} ||q_i(a)||_Q$$

Clearly, $\Delta > 0$. Thus

$$\begin{aligned} ||q_i(a+\gamma c)||_Q &\leq ||q_i(a)||_Q + ||\frac{\theta_i \gamma \alpha_i^T c}{b_i}||_Q \\ &\leq ||q(a)||_Q \end{aligned}$$

provided that we chose γ such that $||\frac{\theta_i \gamma \alpha_i^T c}{b_i}||_Q < \Delta$.

Case 2.1: For $i \in \bar{I}(a)$ in case $\alpha_i a / b_i < 1$, which implies $\theta_i = 1$, we have

$$\begin{aligned} ||q_i(a+\gamma c)||_Q &= ||\frac{\alpha_i^T a}{b_i} + \frac{\gamma \alpha_i^T c}{b_i}||_Q \\ &< ||\frac{\alpha_i^T a}{b_i}||_Q \\ &= ||q_i(a)||_Q \\ &= ||q(a)||_Q \end{aligned}$$

provided that we chose γ such that $0 < \frac{\gamma \alpha_i^T c}{b_i} < 1 - \frac{\alpha_i^T a}{b_i}$.

Case 2.2: For $i \in \bar{I}(a)$ in case $\alpha_i a / b_i \geq 1$, which implies $\theta_i = -1$, we have

$$\begin{aligned} ||q_i(a+\gamma c)||_Q &= ||\frac{\alpha_i^T a}{b_i} + \frac{\gamma \alpha_i^T c}{b_i}||_Q \\ &< ||q_i(a)||_Q \\ &= ||q(a)||_Q \end{aligned}$$

provided that we chose γ such that $\frac{\theta_i \gamma \alpha_i^T c}{b_i} < \frac{\alpha_i^T a}{b_i} - 1$.

Chosing γ small enough such that it satisfies the conditions for all i, we have constructed a contradiction. Thus, $\vec{0} \in D$. By the definition of convex hull, it follows that there exist $I \subseteq \bar{I}$ and γ_i with $\gamma_i > 0$, $\sum_{i \in I} \gamma_i = 1$ and

$$\sum_{i \in \bar{I}} \gamma_i \theta_i \alpha_i = \vec{0}$$

Define

$$\lambda_i = \begin{cases} 0 & i \notin I \\ \gamma_i \theta_i & i \in I \end{cases}$$

Then I and λ_i satisfy the conditions 1-3 of the theorem.

The restriction $|I| \leq n+1$ has not been discussed here, but will follow from the the following lemma. □

Let us illustrate the theorem by our example of section 1.3. For

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}$$

we solve $A^T\lambda = \vec{0}$:

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \vec{0}$$

This results in the (underspecified) system of linear equations

$$\begin{aligned} \lambda_1 + \lambda_2 + \lambda_3 &= 0 \\ \lambda_1 + 2\lambda_2 + 3\lambda_3 &= 0 \end{aligned}$$

Hence, we need $\lambda_2 + 2\lambda_3 = 0$ and thus $\lambda_2 = -2\lambda_3$. Choosing $\lambda_3 = 1$ gives us $\lambda = (1, -2, 1)^T$.

We continue to find bounds on the size of I.

Lemma 7 *If $\sum_{i\in I} \lambda_i\alpha_i = \vec{0}$ and $|I| > n+1$, we can find a non-empty subset I' of I with at most $n+1$ elements and new λ'_i with the same sign as the λ_i such that $\sum_{i\in I'} \lambda'_i\alpha_i = \vec{0}$.*

Proof: Since there are more than $n+1$ elements in I, the vectors α_i, $i \in J$, are linearly dependent for any true subset J of I with at least $n+1$ elements. Thus, there exist γ_i such that

$$\sum_{i\in J} \gamma_i\alpha_i = \vec{0}$$

and, hence, for all δ

$$\delta\sum_{i\in J} \gamma_i\alpha_i = \vec{0}$$

Define $\gamma_i = 0$ for $i \in I \setminus J$, which is non-empty. Then

$$\delta\sum_{i\in I} \gamma_i\alpha_i = \vec{0}$$

Hence,

$$\begin{aligned}\sum_{i\in I}\lambda_i\alpha_i - \delta\sum_{i\in I}\gamma_i\alpha_i &= \vec{0}\\ \sum_{i\in I}(\lambda_i - \delta\gamma_i)\alpha_i &= \vec{0}\end{aligned}$$

Turn $|\delta|$ away from zero until the first λ_k equals $\delta\gamma_k$. Denote by K the set of indices such that λ_k equals $\delta\gamma_k$. Then

$$\sum_{i\in I\backslash K}(\lambda_i - \delta\gamma_i)\alpha_i = \vec{0}$$

and we can compose a new vector λ' from $\lambda'_i = \lambda_i - \delta\gamma_i$ for $i \in I\backslash K$ and $\lambda'_i = 0$ else. Further, $\text{sign}_{\lambda'} = \text{sign}_{\lambda}$ for $i \in I\backslash K$ and λ' is non-trivial by construction. If $I \setminus K$ has more than $n+1$ elements, we repeat the construction. □

An immediate consequence of the theorem and the proof of the lemma is the following:

Corollary 1 *Let a solve 1. Then a solves the problem in R^{n+1} obtained by restricting the components of r to a particular n+1. Further, if A has rank t, then we can restrict the problem to a particular t+1 components.*

In the next subsection, we will show that under a certain condition the solution to our problem is unique. The next lemma and the following corollary are a first step towards this direction. Without any additional conditions, we show in a first step that any solution agrees on the θ_i, i.e., the signs of the residuals. Then, we show that for any two solutions, their residual and quotient vectors are the same.

Lemma 8 *Let a be a solution to Problem 1. Let c be any other solution to Problem 1. Then $\theta_i(a) = \theta_i(c)$ for all $i \in \bar{I}(a)$.*

Proof:

Assume there exists $i \in \bar{I}(a)$ such that $\theta_i(a) \neq \theta_i(c)$. For the following cases we construct a contradiction.

Case 1: If for some $i \in \bar{I}(a)$ we have $\theta_i(a) = 0$ then we must have that for all $i \in \bar{I}(a)$ $\theta_i(a) = 0$. Thus for all $\imath \in \bar{I}(a)$ $\theta_i(c) = 0$ must hold since c is also a solution.

Case 2: $\theta_i(a) < 0$, $\theta_i(c) \geq 0$:

From the fact that a and c are solutions and the definition of $\bar{I}(a)$, we can conclude that $||q_i(a)||_Q \geq ||q_i(c)||_Q$. Thus

$$\begin{array}{rrcl} & ||q_i(a)||_Q & \geq & ||q_i(c)||_Q \\ \iff & \frac{b_i - r_i(a)}{b_i} & \geq & \frac{b_i}{b_i - r_i(c)} \\ \iff & (b_i - r_i(a))(b_i - r_i(c)) & \geq & b_i^2 \end{array}$$

However, Observation 8 d gives us $|r_i(a)| > |r_i(c)|$ and thus

$$\begin{array}{rrcr} & |r_i(a)| & > & |r_i(c)| \\ \iff & (b_i - r_i(a)) & < & (b_i + r_i(c)) \\ \iff & (b_i - r_i(a))(b_i - r_i(c)) & < & b_i^2 - r_i(d)^2 \\ \iff & (b_i - r_i(a))(b_i - r_i(c)) & < & b_i^2 \end{array}$$

We constructed the required contradiction.
Case 3: $\theta_i(a) > 0$, $\theta_i(c) \leq 0$:

From the fact that a and c are solutions and the definition of $\bar{I}(a)$, we can conclude that $||q_i(a)||_Q \geq ||q_i(c)||_Q$. Thus

$$\begin{array}{rrcr} & ||q_i(a)||_Q & \geq & ||q_i(c)||_Q \\ \iff & \frac{b_i}{b_i - r_i(a)} & \geq & \frac{b_i - r_i(c)}{b_i} \\ \iff & (b_i - r_i(a))(b_i - r_i(c)) & \leq & b_i^2 \end{array}$$

However, Observation 8 c gives us $|r_i(a)| < |r_i(c)|$ and thus

$$\begin{array}{rcl} (b_i - r_i(a))(b_i - r_i(c)) & = & b_i^2 - r_i(a)b_i - r_i(c)b_i + r_i(a)r_i(d) \\ & > & b_i^2 + r_i(c)b_i - r_i(c)b_i + r_i(a)r_i(d) \\ & = & b_i^2 r_i(a)r_i(d) \\ & > & b_i^2 \end{array}$$

We constructed the required contradiction.

□

With the help of this lemma, we can proof the following corollary to theorem 5.

Corollary 2 *Let a be a solution to Problem 1. Further chose $I \subseteq \bar{I}(a)$ according to Theorem 5. Let d be any other solution of 1. Then $r_i(a) = r_i(d)$ and $q_i(a) = q_i(d)$ for all $i \in I$.*

Proof: We have

$$\begin{aligned}
\sum_{i\in I}|\lambda_i|\;|1-(||q(a)||_Q)^{-\theta_i(a)}|\;|b_i| &= \sum_{i\in I}|\lambda_i|\;|1-(||q_i(a)||_Q)^{-\theta_i(a)}|\;|b_i| \\
&= \sum_{i\in I}|\lambda_i r_i(a)| \\
&= \sum_{i\in I}\lambda_i r_i(a) \\
&= |\sum_{i\in I}\lambda_i r_i(a)| \\
&= |\sum_{i\in I}\lambda_i(b_i-\alpha_i^T a)| \\
&= |\sum_{i\in I}\lambda_i b_i| \\
&= |\sum_{i\in I}\lambda_i(b_i-\alpha_i^T d)| \\
&= |\sum_{i\in I}\lambda_i r_i(d)| \\
&\leq \sum_{i\in I}|\lambda_i|\;|r_i(d)| \\
&\leq \sum_{i\in I}|\lambda_i|\;|1-(||q_i(d)||_Q)^{-\theta_i(d)}|\;|b_i| \\
&\leq \sum_{i\in I}|\lambda_i|\;|1-(||q(a)||_Q)^{-\theta_i(d)}|\;|b_i| \quad (*)
\end{aligned}$$

To see (*) note that for all $x<y$, $x\geq 1$, $y\geq 1$ we have $\frac{1}{x}>\frac{1}{y}$ and thus

$$\begin{aligned}
|1-x| = x-1 &< y-1 = |1-y| \\
|1-\frac{1}{x}| = 1-\frac{1}{x} &< 1-\frac{1}{y} = |1-\frac{1}{y}|
\end{aligned}$$

From Lemma 8 we can infer that $\theta_i(a)=\theta_i(d)$. Thus, equality holds through and the result follows.

□

The next logical question is how big is $\bar{I}(a)$. It is answered by the next theorem.

Theorem 6 *If A has rank t then there exists a solution a of Problem 1 with $|\bar{I}(a)| \geq t+1$.*

Proof: Let a be any solution with $|\bar{I}(a)| < t+1$. Since the corresponding rows of A are linearly dependent (due to Theorem 5), there exists a non-trivial vector $c \in R^n$ such that for all $i \in \bar{I}(a)$

$$\alpha_i^T c = 0 \tag{2}$$

Thus we have

$$\begin{aligned} ||q_i(a+\gamma c)||_Q &= ||q(a)||_Q \quad \text{for } i \in \bar{I}(a) \\ ||q_i(a+\gamma c)||_Q &= ||\frac{\alpha_i^T a - \gamma\alpha_i c}{b_i}||_Q \quad \text{for } i \notin \bar{I}(a) \end{aligned}$$

Rank $A = t$ implies that for some c satisfying 2 there exists a $j \notin \bar{I}(a)$ with

$$\alpha_i^T c = \delta \neq 0$$

Thus, we can increase $|\gamma|$ away from zero until the first index not in $\bar{I}(a)$ is such that

$$\begin{aligned} ||q_i(a+\gamma c)||_Q &= ||\frac{\alpha_i^T a - \gamma\alpha_i c}{b_i}||_Q \\ &= ||q(a)||_Q \end{aligned}$$

Thus $|\bar{I}(a+\gamma c)| > |\bar{I}(a)|$. We can repeat the process and the claim follows. □

An immediate consequence is the following corollary.

Corollary 3 *The submatrix of A consisting of the rows of A corresponding to $\bar{I}(a)$ must have rank t for some solution a.*

This shows that we essentially have to consider subsets of $t+1$ rows of A to find some $\bar{I}(a)$. This subset is then called an *extremal subset*. It can be characterized as follows.

Theorem 7 *Let A have rank t and let $J \subseteq \{1,\ldots,m\}$ with $|J| = t+1$. Then*

$$\min_a\{\max_{i\in J} ||q_i(a)||_Q\} \leq min_a ||q(a)||_Q$$

Equality holds if and only if J is an extremal subset, i.e. corresponds to a solution of Theorem 5.

Proof: Let q be $q = min_a ||q(a)||_Q$ and d be any solution to Problem 1. Then, for all $i \in \{1, \ldots, m\}$ we have

$$||q_i(d)||_Q \leq q$$

and hence

$$\max_{i \in J} ||q_i(d)||_Q \leq q.$$

The first part of the claim follows from Cor. 1. If J is an extremal subset, equality holds as then $||q_i(d)||_Q = q$ for all $i \in J$, since $J \subseteq \bar{I}(a)$. The other direction also follows from Cor. 1. □

3.6 Uniqueness of Solution

Looking at theorem 5 and corollary 2 it becomes clear that what remains to be shown is that

$$r_i(a) = r_i(d)$$

or

$$q_i(a) = q_i(d)$$

or, equivalently,

$$\alpha_i^T(a - d) = 0$$

implies $a = d$. This is ensured if A satisfies the Haar condition defined next.

Definition 10 *Let A be an $m \times n$ Matrix with $m \geq n$. A satisfies the* Haar condition *if every $n \times n$ submatrix of A is non-singular.*

This means that n arbitrary rows of A are linearily independent. Obviously, this is sufficient for what we have to show.

Theorem 8 *If A satisfies the Haar condition, the solution to Problem 1 is unique.*

Proof: Let a be a solution to Problem 1. Then, by Theorem 5, there exists $I \subseteq \bar{I}(a)$ and a non-trivial vector $\lambda \in R^m$ such that

$$\sum_{i \in I} \lambda_i \alpha_i = \vec{0}.$$

If A satisfies the Haar condition, I must contain at least $n+1$ elements. Let d be any other solution. Then Corollary 2 gives us

$$\alpha_i^T(a-d)=0$$

for all $i \in I$, which is a contradiction to the Haar condition except if $a=d$ which had to be shown.

□

Our goal is to optimally (under the Q-paranorm) approximate histogram bucket with entries (x_i, y_i) for $1 \leq i \leq m$ by polynomials. The vector b is formed by the frequencies y_i. If we want to approximate the bucket by a polynomial of degree k, A becomes the $m \times (k+1)$ Matrix

$$A = \begin{pmatrix} 1 & x_1^1 & x_1^2 & \dots & x_1^k \\ 1 & x_2^1 & x_2^2 & \dots & x_2^k \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_m^1 & x_m^2 & \dots & x_m^k \end{pmatrix}$$

which clearly satisfies the Haar condition.

Although polynomials seem sufficient, our theoretical results allow for more general approximations using any set of continous functions that forms a Chebyshev set (see below).

The following theorem and its proof provide a first good hint for the general outline of the algorithm to come.

Theorem 9 *Let A satisfy the Haar condition and let $J \subset \{1, \dots, m\}$ with $|J| = n+1$. Then, unless J is extremal, it is possible to exchange one index of J to form a new subset J^* such that*

$$q = min_a max_{i \in J} ||q_i(a)||_Q < min_a max_{i \in J^*} ||q_i(a)||_Q = q^*$$

Proof: If J is extremal, no such exchange is possible, since J identifies an optimal solution.

Let a_J be such that a_J minimizes $max_{i \in J}||q_i(a)||_Q$. If J is not extremal then there exists a $\bar{j} \in \{1, \dots, m\} \setminus J$ such that

$$||q_{\bar{j}}||_Q > q$$

If several such $\bar{j}$ exist, chose arbitrarily.

Let $j \in J$ be arbitrary and define $J^* = J \backslash j \cup \{\bar{j}\}$. Let a_{J^*} be the solution to $min_a max_{i \in J^*} ||q_i(a)||_Q$. Since A satisfies the Haar condition, both, a_J and a_{J^*} are unique. We observe that

$$||q_{\bar{j}}(a_{J^*})||_Q \leq ||q_{\bar{j}}(a_J)||_Q$$

due to the construction of J^*.

Since

$$||q_{\bar{j}}(a_{J^*})||_Q = ||q_{\bar{j}}(a_J)||_Q$$

implies that J is an extremal subset, we must have

$$||q_{\bar{j}}(a_{J^*})||_Q < ||q_{\bar{j}}(a_J)||_Q.$$

But then $a_J \neq a_{J^*}$ and thus

$$min_a max_{i \in J} ||q_i(a_J)||_Q < min_a max_{i \in J^*} ||q_i(a_{J^*})||_Q$$

□

3.7 Determining the Solution

In principle it would be possible to try all subsets I of $\{x_1, \ldots, x_m\}$ with $|I| = n+1$. This is however a very inefficient procedure. If we know something more about the θ_i of Theorem 5 then it will be possible to derive a much more efficient algorithm. The following definition specifies our needs.

Definition 11 *Let a be a vector in R^n. We say that $r(a)$* alternates *s times, if there exists points $x_{i_1}, \ldots, x_{i_s} \in \{x_1, \ldots, x_m\}$ such that*

$$r_{i_k}(a) = -r_{i_{k+1}}(a)$$

for $1 \leq k < s$. The set $\{_{i_1}, \ldots, x_{i_s}\}$ is called an alternating set *for a.*

The goal is to show that for every solution there exists an alternating set with $n+1$ points and that every a with an alternating set of size $n+1$ is a solution for this set. Knowing this, we can easily justify the algorithm given in the next section.

In order to proof our goal, we need the notion of Chebyshev set.

Definition 12 *Let X be a closed interval of R. A set of continous function $\Phi_1(x), \ldots, \Phi_n(x)$, $\Phi_i : X \rightarrow R$, is called a* Chebyshev set*, if every non-trivial linear combination of these functions has at most $n-1$ zeros in X.*

Obviously, the set of polynomials $\Phi_i(x) = x^{i-1}$, $1 \leq i \leq n$ forms a Chebyshev set.

From now on, we assume that our x_i are ordered, that is $x_1 < \ldots < x_m$. Further, we define $X = [x_1, x_m]$. We also assume that the matrix A of Problem 1 is defined as

$$A = \left(\begin{array}{ccc} \Phi_1(x_1) & \ldots & \Phi_n(x_1) \\ \ldots & \ldots & \ldots \\ \Phi_1(x_m) & \ldots & \Phi_n(x_m) \end{array} \right)$$

where the Φ_i are contineous functions from X to R. We further assume that they form a Chebyshev set.

The following Lemma is well-known (see [5] page 55). For convenience we also repeat the proof from there.

Lemma 9 *Let $X = [x_1, x_m]$ and let $\Phi_i(x)$, $1 \leq i \leq n$ form a Chebyshev set on X. Let $z_i \in R$ be such that $x_1 \leq z_1 < z_2 < \ldots < z_{n+1} \leq x_m$. Define $A_i(z_1, \ldots, z_i, z_{i+1}, \ldots, z_{n+1})$ as*

$$A_i = \left(\begin{array}{ccc} \Phi_1(z_1) & \ldots & \Phi_n(z_1) \\ \ldots & \ldots & \ldots \\ \Phi_1(z_{i-1}) & \ldots & \Phi_n(z_{i-1}) \\ \Phi_1(z_{i+1}) & \ldots & \Phi_n(z_{i+1}) \\ \ldots & \ldots & \ldots \\ \Phi_1(z_{n+1}) & \ldots & \Phi_n(z_{n+1}) \end{array} \right)$$

and the determinant $\Delta_i = |A_i(z_1, \ldots, z_i, z_{i+1}, z_{n+1})|$. Then sign$(\Delta_i) =$ sign(Δ_{i+1}) for all $1 \leq i \leq n+1$.

Proof: The Chebyshev set assumption implies that $\Delta_i \neq 0$ for all $1 \leq i \leq n+1$. Suppose there exist $j, k \in \{1, \ldots, n_1\}$ such that $\Delta_j < 0 < \Delta_k$ where

$$\begin{array}{rcl} \Delta_j & = & |A_j(z_{j,1}, \ldots, z_{j,n})| \\ \Delta_k & = & |A_j(z_{k,1}, \ldots, z_{k,n})| \end{array}$$

with $z_{j,i} < z_{j,i+1}$ and $z_{k,i} < z_{k,i+1}$ for all $1 \leq i < n$.

Since the Φ_i are continous functions, there must exist $\gamma \in]0,1[$ such that

$$\Delta(\gamma z_{j,1} + (1-\gamma)z_{k,1}, \ldots, \gamma z_{j,n} + (1-\gamma)z_{k,n}) = 0$$

It follows that for some o, p

$$\gamma z_{j,o} + (1-\gamma)z_{k,o} = \gamma z_{j,p} + (1-\gamma)z_{k,p}$$

and, hence,

$$\gamma(z_{j,o} - \gamma z_{j,p}) = (1-\gamma)(z_{k,p} - z_{k,o})$$

which contradicts the ordering assumption. □

Theorem 10 *A vector $a \in R^n$ solves Problem 1 if there exists an alternating set with $n+1$ points for a.*

Proof: Let A fulfill the Haar condition and let a be the solution. Then Theorem 5 gives us $I \subseteq \bar{I}(a)$, $|I| = n+1$, and $\lambda \in R^{n+1}$ with

1. $A^T\lambda = \vec{0}$
2. $\lambda_i\theta_i > 0$ for all $i \in I$
3. $\lambda_i = 0$ for $i \notin I$

Let $x_1, \ldots, x_{n+1}$ be the elements of I. We now need to define several vectors and matrices:

$$\begin{array}{rcl} c_i & = & (\Phi_1(x_i), \ldots, \Phi_n(x_i))^T \\ \lambda^k & = & (\lambda_1, \ldots, \lambda_{k-1}, \lambda_{k+1}, \ldots, \lambda_{n+1})^T \\ B^k & = & (c_1, \ldots, c_{k-1}, c_{k+1}, \ldots, c_{n+1}) \end{array}$$

1. from above implies that for all k

$$B^k\lambda^k = -\lambda_k c_k$$

For $j \neq k$, define $B^{j\leftarrow k}$ depending on $j < k$ or $j > k$:

$$\begin{array}{rcl} B^{j\leftarrow k} & = & (c_1, \ldots, c_{j-1}, \lambda_k c_k, c_{j+1} \ldots, c_{k-1}, \ldots, c_{k+1}, \ldots, c_{n+1})^T \\ B^{j\leftarrow k} & = & (c_1, \ldots, c_{k-1}, \ldots, c_{k+1}, \ldots, c_{j-1}, \lambda_k c_k, c_{j+1}, \ldots, c_{n+1})^T \end{array}$$

Further note that we need exactly $|j-k|+1$ column exchanges to go from $B^{j \leftarrow k}$ to B^j. Cramer's rule gives us for all $j \neq k$:

$$\begin{aligned} \lambda_j &= (-1)\lambda_k \frac{detB^{j \leftarrow k}}{detB^k} \\ &= (-1)^{|j-k|}\lambda_k \frac{detB^j}{detB^k} \end{aligned}$$

Since Lemma 9 implies that $\frac{detB^j}{detB^k} > 0$ the claim follows.

□

As we can see from the proof, the only property that is really necessary is that the determinants of the A_i in Lemma 9 have the same sign. Thus, we can generalize our results beyond Chebyshev sets to any set of functions with this property. This allows us to use non-continous functions which might be useful in some situations.

The other direction of Theorem 10 is not valid as for every set I with $|I| = n+1$ there exists a unique alternate solution. However, we have the following theorem.

Theorem 11 *Let I be a subset of $\{1, \dots, m\}$ with $|I| = n+1$ and $a \in R^n$ such that I is an alternating set for a. Then a is optimal for I, i.e.*

$$\max_{i \in I} ||q_i(a)||_Q = \min_c \max_{i \in I} ||q_i(c)||_Q.$$

Proof: Assume there exists an optimal solution for $c \in R^n$ for I and

$$\max_{i \in I} ||q_i(a)||_Q > \max_{i \in I} ||q_i(c)||_Q.$$

Theorem 10 implies that I is an alternating set for c. Consider the following two cases.

In case 1, $\theta_i(r_i(a)) = \theta_i(r_i(c))$ for all $i \in I$. Then, we have that

$$\begin{aligned} \sum_{j=1}^{n} (a_j - c_j)\Phi_j(x_i) &= \alpha_i^T(a_j - c_j) \\ &= \alpha_i^T a_j - \alpha_i^T c_j \\ &= (b_i - \alpha_i^T c) - (b_i - \alpha_i^T a) \\ &= r_i(c) - r_i(a) \end{aligned}$$

alternates in sign as i goes through I. As the Φ_i are continous, there must exist n points $z_1, \ldots, z_n$ in X such that

$$\sum_{j=1}^{n} (a_j - c_j)\Phi_j(z_i) = 0$$

for all $1 \leq i \leq n$ which contradicts the Chebyshev set assumption.

In case 2, $\theta_i(r_i(a)) = -\theta_i(r_i(c))$ for all $i \in I$. We can construct a solution $d = \frac{a+b}{2}$ with

$$\max_{i \in I} ||q_i(d)||_Q < \max_{i \in I} ||q_i(c)||_Q$$

which is a contradiction to the optimality of c.

□

4 Approximation by Linear Functions: $n = 2$

4.1 Q-paranorm

The above theorems assure us the uniqueness and existence of a solution. Moreover, for every subset I of indices with $|I| = n + 1$, there exists an alternating set. This allows us to derive a solution for a particular $I = \{x_1, x_2, x_3\}$ by solving the following system of three equations:

$$\begin{array}{rcl} \frac{1}{\lambda}(\alpha + \beta x_1) & = & y_1 \\ \lambda(\alpha + \beta x_2) & = & y_2 \\ \frac{1}{\lambda}(\alpha + \beta x_3) & = & y_3 \end{array}$$

$$\begin{array}{lrcl} 3 \Longrightarrow & \alpha & = & \lambda y_3 - \beta x_3 \quad (*) \\ 1, (*) \Longrightarrow & \lambda y_3 - \beta x_3 + \beta x_1 & = & \lambda y_1 \\ \Longrightarrow & (y_3 - y_1)\lambda & = & (x_3 - x_1)\beta \\ \Longrightarrow & \lambda & = & \frac{x_3 - x_1}{y_3 - y_1}\beta \quad (**) \\ \Longrightarrow & \lambda & = & q_{13}\beta \quad (**) \\ 2, (*), (**) \Longrightarrow & q_{13}\beta(q_{13}y_3\beta - \beta x_3 + \beta x_2) & = & y_2 \\ \Longrightarrow & \beta^2(q_{13}y_3 - x_3 + x_2) & = & y_2 q_{13}^{-1} \\ \Longrightarrow & \beta & = & \sqrt{g^{-1} y_2 q_{13}^{-1}} \end{array}$$

where

$$\begin{aligned} q_{13} &:= \frac{x_3 - x_1}{y_3 - y_1} \\ g &:= q_{13}y_3 - x_3 + x_2 \end{aligned}$$

Caution is necessary, if $\beta = 0$. Then:

$$\begin{aligned} \beta &= 0 \\ \alpha &= \lambda y_1 \\ \lambda &= \sqrt{y_2/y_1} \end{aligned}$$

4.2 From Chebyshev to Q

The following approach often gives very good approximations. Let (x_i, y_i) be the data set we want to approximate. Instead of approximating this directly, we approximate $(x_i, \ln y_i)$ with the Chebyshev norm. The approximation function then is e^{b+ax} and the norm minimized on the orignal data is the Q-paranorm

4.3 From Q to Chebyshev

We can also perform the dual of the previous subsection. Let (x_i, y_i) be the data we want to approximate by a function of the form $\ln(b + ax)$ while minimizing the Chebyshev norm. We can do so by approximating (x_i, e^{y_i}) by a linear function while minimizing the Q-norm.

5 Algorithm

The algorithm we discuss can be used for any of the (para-) norms mentioned in the introduction. For each of them, a specific system of (linear) equations has to be solved as we did for the Q-paranorm. Generating the solutions is a subroutine of the following algorithm and the only part that is dependent on the (para-) norm used. The appendix gives the equations and solutions thereof for the norms mentioned in the introduction.

5.1 Exchange Rule

For given i_1, i_2, i_3 with $x_{i_1} < x_{i_2} < x_{i_3}$ and derived α, β, λ, we try to find new indices j_1, j_2, j_3 by exchanging one of the i_j with k such that λ will be increased. (The λ is dependable on the (para-) norm used.) Assume the deviation of the (current) estimate $\hat{f}$ is maximized at some k. Then, we will exchange one of the i_1, i_2, i_3 by k according to the following exchange rule. Define $\hat{f}_i = \alpha + \beta x_i$. Depending on the position of x_k in the sequence i_1, i_2, i_3 and the sign of the residual we determine the i_j to be exchanged with k:

- $x_k < x_{i_1}$
  if $(\text{sign}(y_k - \hat{f}_k) == \text{sign}(y_{i_1} - \hat{f}_{i_1}))$
  then $j_1 = k, j_2 = i_2, j_3 = i_3$
  else $j_1 = k, j_2 = i_1, j_3 = i_2$

- $x_{i_1} < x_k < x_{i_2}$
  if $(\text{sign}(y_k - \hat{f}_k) == \text{sign}(y_{i_1} - \hat{f}_{i_1}))$
  then $j_1 = k, j_2 = i_2, j_3 = i_3$
  else $j_1 = i_1, j_2 = k, j_3 = i_2$

- $x_{i_2} < x_k < x_{i_3}$
  if $(\text{sign}(y_k - \hat{f}_k) == \text{sign}(y_{i_2} - \hat{f}_{i_2}))$
  then $j_1 = i_1, j_2 = k, j_3 = i_2$
  else $j_1 = i_1, j_2 = i_2, j_3 = k$

- $x_k > x_{i_3}$
  if $(\text{sign}(y_k - \hat{f}_k) == \text{sign}(y_{i_3} - \hat{f}_{i_3}))$
  then $j_1 = i_1, j_2 = i_2, j_3 = k$
  else $j_1 = i_2, j_2 = i_3, j_3 = k$

5.2 Algorithm

1. Choose arbitrary i_1, i_2, i_3 with $x_{i_1} < x_{i_2} < x_{i_3}$.
   (In our implementation we used equi-distant i_j.)

2. Calculate the solution for the system of equations corresponding to the (para-) norm used.
   This gives us an approximation function $\hat{f}(x) = \alpha + \beta x$ and λ.

3. Find an x_k for which the deviation of $\hat{f}$ from the given data is maximized. Call this maximal deviation maximum $\lambda_{\max}$.

4. If $\lambda_{\max} - \lambda > \epsilon$ for some small ϵ
   then apply the exchange rule using x_k and $\lambda_{\max}$ and go to step 2.
   (The ϵ is mainly needed for rounding problems with floating point numbers. If they were non-existant, one could choose $\lambda_{\max} \neq \lambda$ as the criterion.)

5. Return α, β, λ.

6 Practical Tips to Improve the Accuracy

6.1 Intervals

The approach we presented optimizes the maximum error for single-point queries. In order to see the problems occurring with real data when estimating frequency counts for ranges, let us take a look at an example. Fig 1 contains the number of authors for a given number of citations as extracted from the citesee top 10.000 cited computer science authors. The figure only shows the number of authors cited between 256 and 512 times.

Remember that the Q-paranorm is a number always larger than 1 and we cannot see from the Q-paranorm whether we have an underestimate or an overestimate. We thus define the Q-error for visualization purposes as follows:

$$\theta_i(||q_i||_Q - 1)$$

Let us now turn to range queries. For a given interval length, we can calculate the minimum, average, and maximum Q-error over all possible ranges of this length. The upper part of Fig. 2 visualizes the result. We observe that the minimum and maximum Q-error converge to the average quite fast. We further observe that the average is above 0. This means, that on the average the estimation underestimates. This could also be guessed from looking at Fig. 1. The idea to improve the accuracy for range queries is quite simple: Multiple the result by the average error. Therefore, we approximate the average error by a linear function. For an interval length ≥ 5, we apply the correction. We can then calculate the Q-error of the corrected estimate. This is given at the bottom of Fig. 2. We see that the convergence is quicker and that the average Q-error is 0 for ranges containing at least 5 points.

Another possibility one could think of is to smooth the original histogram with a kernel. For example, we could replace the frequency at each point by

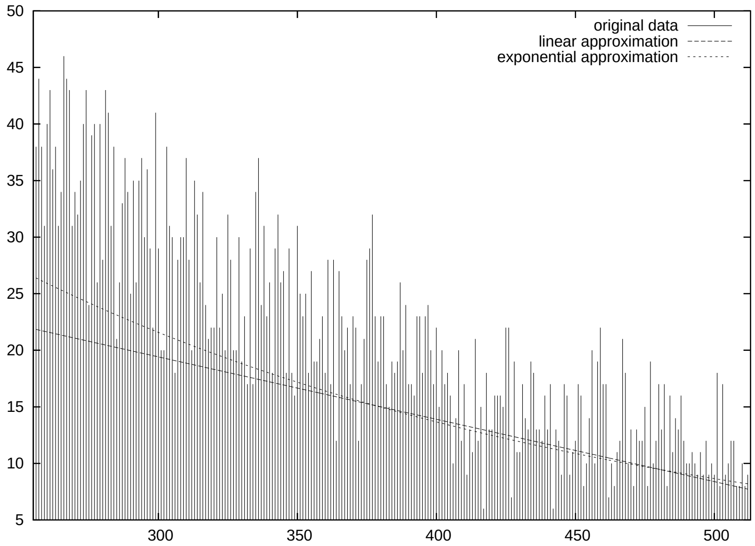

Figure 1: Original Data and Approximations

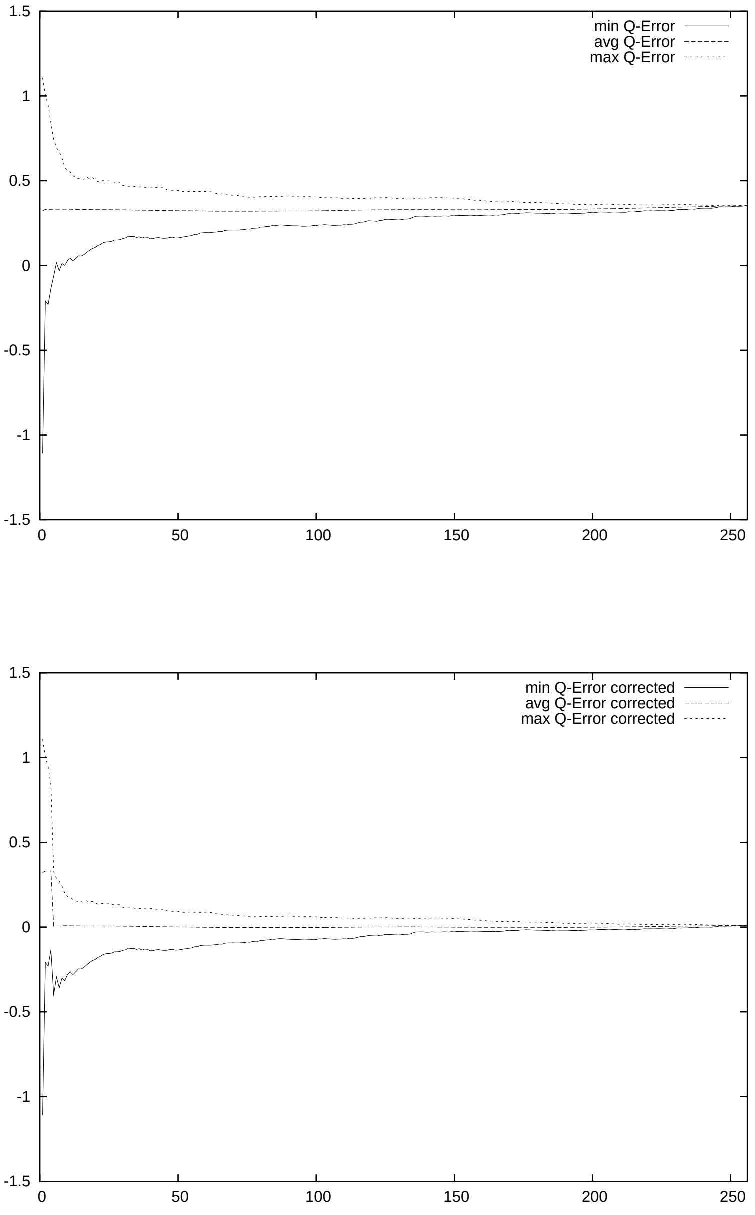

Figure 2: Q-error for a given interval length
The picture shows the minimum, average and maximum Q-error for the original approximation (top) and the corrected approximation (bottom).

the avg frequency of the point itself and four of its neighbors. This smoothed histogram could then be approximated the way we did and it could be hoped that the resulting approximation is better then the original one. This is indeed true, however, the approach is inferior to the one sketched above.

6.2 Single Point

Another possible improvement is to eliminate outliers from being estimated by the approximation. How do we detect the outliers to be removed? This is easy since the algorithm directly constructs three indices where the deviation from the approximation is maximized. Hence, one of these three points should be a good candidate for the outlier. Which one to chose? This is, again, easy. As we saw in the previous section, the average estimation error is typically not zero but greater (as above) or smaller. If we overestimate more often than underestimate, this means that the underestimates are the true outliers, and vice versa. In the above example, the optimal approximation under the Q-paranorm on the average underestimates the true values. Hence, we chose the index as an outlier whose estimation results is an overestimate. The following table shows the Q-paranorm depending on how many outliers have iteratively been removed according to the above rule.

Outliers removed	Q-Norm
0	2.16058
1	2.04246
2	1.95729
3	1.79868
4	1.78647
5	1.72189
6	1.71228
7	1.69124
8	1.67364
9	1.66759
10	1.65571
11	1.64367
12	1.64220
13	1.61936
14	1.60921
15	1.58685
16	1.55443
17	1.55354
18	1.55094
19	1.53623
20	1.52118
21	1.51304
22	1.50971
23	1.50405
24	1.50389
25	1.49071

Note that the above rule is important. Picking any outlier among the three points that maximize the Q-paranorm is no good idea. For example, if we accidentally implement the opposite of the above procedure, after 10 iterations we would still have a Q-paranorm of 2.07862.

Another trick is to keep the list of points where over- or underestimation occurs. If it is possible due to a dense domain, to keep them as a bitmap, which is then compressed, this might be affordable (in terms of memory) and pay off (in terms of gained accuracy).

Last but not least, it is possible to split the bucket to improve the estimation quality. In our case, the optimal split occurs at 401 and the maximum

Q-Norm within the two intervals are

Interval	Q-Norm
[256,401]	1.66141
[402,512]	1.82574

Hence, the overall Q-Norm becomes 1.8. Note that storing only 3 outliers is necessary in order to beat this refinement. If we assume that we need 3 bytes to store each outlier, we need 9 bytes to do so. The additional storage requirement for storing the refinement amounts to 10 bytes. Two bytes for the additional boarder and two times 4 bytes for the additional α and β.

A Approximation by Linear Functions: $n = 2$

A.1 Chebyshev-Norm: $|| \cdot ||_\infty$

Let $r = (r_1, \dots, r_m)^T$ be a vector in R^m. Then, the Chebyshev norm is defined as

$$||r||_C = \max_{1 \le i \le m} |r_i|$$

The problem considered in this section is

$$\text{find } a \in R^n \text{ to minimize } ||r(a)||_C$$

where

$$r(a) = b - Aa$$

with A a given $m \times n$ matrix and b a given vector in R^m.

The solution to this general problem as well as algorithms are described in the book by Watson [5].

In case of a simple linear function used to approximate a set of points subject to minimizing the Chebyshev norm, this amounts to find α, β which minimize

$$\begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \dots \end{pmatrix} - \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \multicolumn{2}{c}{\dots} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

With $b = (f_1, \ldots, f_m)^T$,

$$A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ & \ldots \end{pmatrix}$$

This corresponds to minimizing $b - Aa$.

For simple linear functions, the maximum difference is attained at three points:

Theorem 12 *Let A satisfy the Haar condition. Then, there exist i_1, i_2, i_3 in $\{1, \ldots, m\}$ such that*

$$(-1)^n (y_{i_j} - f_{i_j}) = \max_i |y_i - f_i|$$

for all $1 \leq j \leq 3$ and $1 \leq i \leq m$.

This gives a hint on how to find α, β. We introduce an additional variable λ for $(-1)^n \min_a \max_{j=1,2,3} |y_{i_j} - f_{i_j}|$. Now, we can solve the system of three equations

$$\begin{aligned} -1(y_{i_1} - (\alpha + \beta x_{i_1})) &= \lambda \\ +1(y_{i_2} - (\alpha + \beta x_{i_2})) &= \lambda \\ -1(y_{i_3} - (\alpha + \beta x_{i_3})) &= \lambda \end{aligned}$$

which can be rewritten to

$$\begin{aligned} \alpha + \beta x_{i_1} - \lambda &= y_{i_1} \\ \alpha + \beta x_{i_2} + \lambda &= y_{i_2} \\ \alpha + \beta x_{i_3} - \lambda &= y_{i_3} \end{aligned}$$

or

$$\begin{pmatrix} 1 & x_{i_1} & -1 \\ 1 & x_{i_2} & +1 \\ 1 & x_{i_3} & -1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \lambda \end{pmatrix} = \begin{pmatrix} y_{i_1} \\ y_{i_2} \\ y_{i_3} \end{pmatrix}$$

With

$$\begin{pmatrix} 1 & x_{i_1} & -1 & y_{i_1} \\ 1 & x_{i_2} & +1 & y_{i_2} \\ 1 & x_{i_3} & -1 & y_{i_3} \end{pmatrix} \rightsquigarrow \begin{pmatrix} 1 & x_{i_1} & -1 & y_{i_1} \\ 0 & x_{i_2} - x_{i_1} & +2 & y_{i_2} - y_{i_1} \\ 0 & x_{i_3} - x_{i_1} & 0 & y_{i_3} - y_{i_1} \end{pmatrix} \rightsquigarrow \begin{pmatrix} 1 & x_{i_1} & -1 & y_{i_1} \\ 0 & 1 & \frac{2}{x_{i_2} - x_{i_1}} & \frac{y_{i_2} - y_{i_1}}{x_{i_2} - x_{i_1}} \\ 0 & 1 & 0 & \frac{y_{i_3} - y_{i_1}}{x_{i_3} - x_{i_1}} \end{pmatrix}$$

$$\rightsquigarrow \begin{pmatrix} 1 & x_{i_1} & -1 & y_{i_1} \\ 0 & 1 & \frac{2}{x_{i_2}-x_{i_1}} & \frac{y_{i_2}-y_{i_1}}{x_{i_2}-x_{i_1}} \\ 0 & 0 & \frac{-2}{x_{i_2}-x_{i_1}} & \frac{y_{i_3}-y_{i_1}}{x_{i_3}-x_{i_1}} - \frac{y_{i_2}-y_{i_1}}{x_{i_2}-x_{i_1}} \end{pmatrix} \rightsquigarrow \begin{pmatrix} 1 & x_{i_1} & -1 & y_{i_1} \\ 0 & 1 & \frac{2}{x_{i_2}-x_{i_1}} & \frac{y_{i_2}-y_{i_1}}{x_{i_2}-x_{i_1}} \\ 0 & 0 & 1 & \frac{y_{i_2}-y_{i_1}}{2} - \frac{(y_{i_3}-y_{i_1})(x_{i_2}-x_{i_1})}{2(x_{i_3}-x_{i_1})} \end{pmatrix}$$

we get

$$\begin{aligned} \lambda &= \frac{y_{i_2}-y_{i_1}}{2} - \frac{(y_{i_3}-y_{i_1})(x_{i_2}-x_{i_1})}{2(x_{i_3}-x_{i_1})} \\ \beta &= \frac{y_{i_2}-y_{i_1}}{x_{i_2}-y_{i_1}} - \frac{2\lambda}{x_{i_2}-x_{i_1}} \\ \alpha &= y_{i_1} + \lambda - x_{i_1}\beta \end{aligned}$$

A.2 Relative Differences

A.2.1 Relative Difference: S_R

$$max_i|\frac{y_i - f_i}{y_i}|$$

Gleichungssystem:

$$\begin{aligned} \beta + x_1\alpha &= (1+\lambda)y_1 \\ \beta + x_2\alpha &= (1-\lambda)y_2 \\ \beta + x_3\alpha &= (1+\lambda)y_3 \end{aligned}$$

$$\begin{array}{lrcl} 1 \Longrightarrow & \beta &=& (1+\lambda)y_1 - x_1\alpha \quad (*) \\ 3,(*) \Longrightarrow & (1+\lambda)y_1 - x_1\alpha + x_3\alpha &=& (1+\lambda)y_3 \\ \Longrightarrow & (1+\lambda)(y_1-y_3) &=& (x_1-x_3)\alpha \\ \Longrightarrow & \alpha &=& (1+\lambda)\frac{y_1-y_3}{x_1-x_3} \\ \Longrightarrow & \alpha &=& (1+\lambda)q_{13} \\ 2,(*),(**) \Longrightarrow & (1+\lambda)y_1 - x_1\alpha + x_2\alpha &=& (1-\lambda)y_2 \\ \Longrightarrow & (1+\lambda)y_1 + (x_2 - x1)(1+\lambda)q_{13} &=& (1-\lambda)y_2 \\ \Longrightarrow & (1+\lambda)(y_1 + (x_2 - x1)q_{13}) &=& (1-\lambda)y_2 \\ \Longrightarrow & (1+\lambda)g &=& (1-\lambda)y_2 \\ \Longrightarrow & g + g\lambda &=& y_2 - y_2\lambda \\ \Longrightarrow & \lambda &=& \frac{y_2-g}{y_2+g} \end{array}$$

where

$$\begin{aligned} q_{13} &= \frac{y_1 - y_3}{x_1 - x_3} \\ g &= y_1 + q_{13}(x_2 - x_1) \end{aligned}$$

A.2.2 Relative Difference: S'_R

$$max_i|\frac{y_i - f_i}{f_i}|$$

Gleichungssystem:

$$\begin{aligned} (1+\lambda)(\beta + x_1\alpha) &= y_1 \\ (1-\lambda)(\beta + x_2\alpha) &= y_2 \\ (1+\lambda)(\beta + x_3\alpha) &= y_3 \end{aligned}$$

$$\begin{array}{lrcl}
1 \Longrightarrow & \beta & = & \frac{y_1}{1+\lambda} - x_1\alpha \quad (*) \\
3,(*) \Longrightarrow & (1+\lambda)(\beta + x_3\alpha) & = & y_3 \\
\Longrightarrow & (1+\lambda)[\frac{y_1}{1+\lambda} + \alpha(x_3 - x_1)] & = & y_3 \\
\Longrightarrow & y_1 + (1+\lambda)\alpha(x_3 - x_1) & = & y_3 \\
\Longrightarrow & (1+\lambda)\alpha(x_3 - x_1) & = & y_3 - y_1 \\
\Longrightarrow & (1+\lambda)\alpha & = & \frac{y_3 - y_1}{x_3 - x_1} \\
\Longrightarrow & \alpha & = & \frac{1}{1+\lambda} q_{13} \quad (**) \\
2,(*),(**) \Longrightarrow & (1-\lambda)(\frac{y_1}{1+\lambda} - x_1\frac{1}{1+\lambda}q_{13} + \frac{1}{1+\lambda}q_{13}x_2) & = & y_2 \\
\Longrightarrow & (1-\lambda)(y_1 - x_1 q_{13} + q_{13}x_2) & = & (1+\lambda)y_2 \\
\Longrightarrow & (1-\lambda)g & = & (1+\lambda)y_2 \\
\Longrightarrow & g - \lambda g & = & y_2 + \lambda y_2 \\
\Longrightarrow & \lambda & = & \frac{g - y_2}{g + y_2} \quad (***)
\end{array}$$

where

$$\begin{aligned} q_{13} &= \frac{y_3 - y_1}{x_3 - x_1} \\ g &= y_1 + q_{13}(x_2 - x_1) \end{aligned}$$

B Proof of the Original Corollary 2 of [5]

Corollary 4 (Cor. 2 of [5]) *Let a solve (2.1) and let $I \subseteq \bar{I}(a)$ be such that $\lambda_i \neq 0$ for all $i \in I$ and $\sum_{i \in I} \lambda_i \alpha_i = \vec{0}$ according to the theorem. Then $r_i(d) = r_i(a)$ for all solutions d of (2.1).*

Proof: Let $h = ||r(a)||$, and let d be any other solution to (2.1). If $h = 0$ the result is trivial, so assume that $h > 0$. Then by the theorem (2):

$$\sum_{i \in I} \lambda_i \alpha_i = \vec{0}$$

and $\lambda_i \theta_i > 0$ for all $i \in I$. Thus:

$$\begin{aligned} h \sum_{i \in I} |\lambda_i| &= \sum_{i \in I} \theta_i \lambda_i \ \theta_i r_i(a) \\ &= \sum_{i \in I} \lambda_i r_i(a) \\ &= |\sum_{i \in I} \lambda_i r_i(a)| \\ &= |\sum_{i \in I} \lambda_i (b_i - \alpha_i^T a)| \\ &= |\sum_{i \in I} \lambda_i b_i| \\ &= |\sum_{i \in I} \lambda_i (b_i - \alpha_i^T d)| \\ &= |\sum_{i \in I} \lambda_i r_i(d)| \\ &\leq \sum_{i \in I} |\lambda_i| \ |r_i(d)| \\ &\leq h \sum_{i \in I} |\lambda_i| \end{aligned}$$

and equality holds through. Since no $|r_i(d)|$ can be larger than h [d is solution!] if follows from

$$\sum_{i \in I} |\lambda_i| \ |r_i(d)| = h \sum_{i \in I} |\lambda_i|$$

and the definition of h and I that $|r_i(d)| = h = |r_i(a)|$.

It remains to be shown that $\theta_i(a) = \theta_i(d)$. From

$$|\sum_{i \in I} \lambda_i r_i(d)| = \sum_{i \in I} \ |\lambda_i| \ |r_i(d)|$$

it follows that

$$\begin{array}{ll} (a) & (\forall i \in I \ \text{sign}(\lambda_i r_i(d)) = -1) \\ & \vee \\ (b) & (\forall i \in I \ \text{sign}(\lambda_i r_i(d)) = +1) \end{array}$$

Assume (a) holds. (a) implies for all $i \in I$ that $\theta_i(a) = -\theta_i(d)$ and $r_i(a) = -r_i(d)$. Summing both sides of

$$\begin{array}{rcl} r(a) & = & b - Aa \\ -r(a) & = & b - Ad \end{array}$$

gives

$$\begin{array}{rcl} \vec{0} & = & 2b - A(a-d) \\ & = & b - 1/2(A(a-d) \\ & = & b - A(1/2(a-d)) \end{array}$$

which contradicts the optimality of a. Thus (b) must hold and the result follows. □

References

[1] P. Gibbons. Distinct sampling for highly-accurate answers to distinct values queries and event reports. In *Proc. Int. Conf. on Very Large Data Bases (VLDB)*, pages 541–550, 2001.

[2] Y. E. Ioannidis and S. Christodoulakis. On the propagation of errors in the size of join results. In *Proc. of the ACM SIGMOD Conf. on Management of Data*, pages 268–277, 1991.

[3] A. König and G. Weikum. Combining histograms and parametric curve fitting for feedback-driven query result-size estimation. In *Proc. Int. Conf. on Very Large Data Bases (VLDB)*, pages 423–434, 1999.

[4] M.Charikar, S. Chaudhuri, R. Motwani, and V. Narasayya. Towards estimation error guarantees for distinct values. In *Proc. ACM SIGMOD/SIGACT Conf. on Princ. of Database Syst. (PODS)*, pages 268–279, 2000.

[5] G. Watson. *Approximation Theory and Numerical Methods.* Addison-Wesley, 1980.