Estimating L1-Norm Best-Fit Lines for Data

J. Paul Brooks[a,b], José H. Dulá[b,*]

[a] *Department of Statistical Sciences and Operations Research,*
[b] *Department of Supply Chain Management and Analytics,*
Virginia Commonwealth University, Richmond, VA 23284

Abstract

The general formulation for finding the L1-norm best-fit subspace for a point set in m-dimensions is a nonlinear, nonconvex, nonsmooth optimization problem. In this paper we present a procedure to estimate the L1-norm best-fit one-dimensional subspace (a line through the origin) to data in $\Re^m$ based on an optimization criterion involving linear programming but which can be performed using simple ratios and sortings. The procedure has distinct advantages in that it does not depend on any initializations, is deterministic and replicable, and is scalable. The estimated line is sharp in that it satisfies a well-defined optimization criterion, and is often tight, meaning that it is sometimes a globally optimal solution. We show how the method can be extended to a procedure for an L1-norm principal component analysis by iteratively approximating higher-order best-fit subspaces. In a comprehensive computational study involving synthetic and real data, the procedure is shown to be more robust to outlier observations than competing approaches.

Keywords: Analytics, L1-norm, line location, principal component analysis
2010 MSC: 90-C05, 90-B85, 62-H30

1. Introduction

Many decision-support analytics models rely upon fitting subspaces to data. Examples include linear regression, logistic regression, and principal component

*Corresponding author
Email address: jdula@vcu.edu (José H. Dulá)

analysis (PCA). Linear and logistic regression fit hyperplanes to data, which are $(m-1)$-dimensional translated subspaces for m-dimensional data. A hyperplane that best fits data by minimizing the sum of squares of projection distances is the foundation of orthogonal regression and fitting all lower dimensional subspaces using this criterion is the essence of PCA [1, 2]. Replacing the sum of squares criterion with the L1 norm results in analogous tools. The properties of the L1 norm make the models robust and desirable in many settings such as when dealing with outliers and noisy data. The best-fit hyperplane for an m-dimensional point set under the L1 norm can be found efficiently by solving m linear programs (LPs) [3, 4]. The best-fit hyperplane turns out to be a special case and a workable solution is not available for subspaces with fewer nonzero dimensions; indeed there is evidence that in those cases, the problem becomes intractable. The topic of this work is the estimation of an L1-norm best-fit one-dimensional subspace for multivariate data, i.e., a line through the origin for a point set in m dimensions. We present an approach with modest computational requirements, that obviates specifying initial starting points, and with superior performance compared to competing alternatives. Moreover, it can generate the best-fit line in some cases.

Suppose we are given points $x_i \in \Re^m$, $i = 1, \ldots, n$. An L1-norm best-fit line through the origin can be found by solving the following unconstrained optimization problem:

$$\min_{\substack{v \in \Re^m, \alpha_i \in \Re; \\ i=1,\ldots,n}} \sum_{i=1}^{n} \|x_i - v\alpha_i\|_1. \tag{1}$$

At optimality, v^* is the direction of an L1-norm best-fit line through the origin and for each point x_i, the optimal coefficient α_i^* indicates the location of an L1 projection of x_i on the line defined by v^*. The problem in (1) is nonlinear, nonconvex, and nondifferentiable with a potential for a multitude of local optima.

The optimization problem presented in (1) is one of many instances of a more general problem of fitting an affine set (translated subspace) to a point set. The fitted object can be expressed analytically and is such that the sum

of the points' distances to it are minimized in either a fixed or free direction and according to a specified norm. The first occurrence of this problem is perhaps the Fermat-Torricelli Problem dating to the 17th century [5, 6]. Ordinary least squares regression is an example where the fitted affine set is a hyperplane and distances are measured parallel to a fixed "response" axis using the sum of squares criterion. Another example is orthogonal regression where a hyperplane is also fitted using the same criterion but without restricting the direction for the points' projections. In many of these manifestations, the objective is to replace the data represented by a point set with an affine set that best summarizes and describes information about the data's location, orientation, and dispersion. These problems are known by different names and appear in different areas including mathematics, statistics, and computer science with all sorts of applications.

There are two types of studies that can be cited as background for the L1-norm best-fit line problem. One is work that specifically treats fitting lines to point sets using the L1 norm, with applications in location theory and ranking. The other is the more general subspace estimation with the L1 norm studied for its own sake or as a subproblem in methodologies such as PCA, low-rank matrix approximation, and low-distortion subspace embedding.

The L1-norm best-fit line problem for data in two-dimensions was treated by Megiddo and Tamir [7], who show that the problem is optimally solved by sorting the ratios of the two coordinates. A line in two-dimensions is a hyperplane so this is the special case of the L1-norm best-fit hyperplane problem in general dimensions for which an efficient solution is known [3, 4]. Heuristic and exponential-time global optimization procedures have been proposed for data in three dimensions [8, 9]. Robust estimates rely on a good initial starting point and use a surrogate optimization function (e.g., [10, 11, 12]) or have exponential complexity (e.g., [13]). To the best of our knowledge, the L1-norm best-fit line problem has not been directly attempted beyond three dimensions.

The more general L1-norm best-fit subspace problem is also relevant as a background. Candès et al. and Goldfarb et al. [14, 15] propose solution methods

for optimization problems that simultaneously penalize approximations of the rank and the reconstruction error of a fitted subspace. An active area of research requiring the use of best-fit subspaces of different dimensions is PCA. PCA is based on minimizing projection distances using the sum of squares. Traditional PCA provides best-fit subspaces for each dimension $1, \ldots, m$, each of which is the solution to an optimization problem based on minimizing the projections using the sum of squares. PCA becomes "robust" if the L1 norm becomes involved. In such robust PCA procedures, subspaces are fitted to the data using the L1 norm. Ke and Kanade [16, 17], Tsagkarakis et al. [18], and Jiang et al. [19] consider the problem of finding a subspace such that the sum of L_1 distances of points to the subspace is minimized, and propose heuristic schemes that approximate the subspace. Brooks et al. [20] find successive globally-optimal L1-norm best-fit hyperplanes in polynomial time using LP for a "backwards" PCA [21]. Clarkson et al. [22, 23] demonstrate how the hyperplanes may be estimated quickly using randomized algorithms, and Visentin et al. [24] propose an LP-based approximation for successive best-fit hyperplanes. Park and Klabjan [25] use iteratively reweighted least squares algorithms for estimating a subspace that minimizes the L1-norm distances of points to their L2-norm projections.

The true complexity of the L1-norm best-fit line problem for m-dimensional data has only recently been established. Gillis and Vavasis [26] demonstrated that the L1-norm best-fit line problem for m-dimensional data is NP-hard. Prior to this, this complexity was only speculation; for example Markopoulos et al. [13] proved that a related problem, that of maximizing the sum of L1-norm lengths of the (L2-norm) projections of points onto a line, is NP-hard and provided an $O(n^m)$ exact algorithm for n points.

The nature of the function being minimized in (1) can be appreciated through an example with 110 points in $\Re^3$. Figure 1 depicts the objective function values along with values for selected solutions (corresponding to different fitted lines) appearing as the small colored spheres. The function has no discernible pattern and at least two local minima. A method requiring an initial starting point can converge to a sub-optimal local optimum. Several purportedly robust methods

converge to a poor local minimum, motivating the need for an efficient procedure that retains robustness. The method proposed herein and L1-PCA* [20] appear to provide a global optimal solution for this instance, while all remaining methods either converge to a local minimum or other suboptimal point.

Some notes on notation: vectors are indicated by lowercase letters, matrices are indicated by capital letters, the transpose of a vector x is indicated by x^T. Observations or points are column vectors. The index i indicates the observation or point, and the index j indicates the attribute or variable; if in the context one index is fixed, then the fixed index may be omitted (e.g., x_j). The unit direction e_j is the vector in $\Re^m$ with the j^{th} coordinate having value one, and all other coordinates zero. A vector x with all coordinates zero is written as $x = 0$; if a vector has at least one non-zero coordinate value, we may write $x \neq 0$.

2. Equivalent Formulation

By introducing two goal variables λ_{ij}^+, λ_{ij}^- for each observation-attribute pair, the optimization problem in (1) can be recast as the following constrained mathematical program:

$$\min_{\substack{v\in\Re^m,\alpha\in\Re^n,\\ \lambda^+,\lambda^-\in(\Re^{m\times n})^+}} \sum_{i=1}^{n}\sum_{j=1}^{m}(\lambda_{ij}^+ + \lambda_{ij}^-); \tag{2}$$

subject to:

$$v_j\alpha_i + \lambda_{ij}^+ - \lambda_{ij}^- \quad = \quad x_{ij}, \quad i = 1,\ldots,n,\ j = 1,\ldots,m.$$

Proposition 1. *Formulation* (2) *is equivalent to* (1).

Proof. A proof is in the Appendix. □

The mathematical program in (2) manages to avert the absolute value operation altogether although nonlinearity remains, having been transferred from the objective function to the constraints. Each constraint contains exactly one bilinear term. An optimal solution to (2) will be a vector $v^* \in \Re^m$, along with values α_i^*, one for each of the n data points, and for each point i, pairs

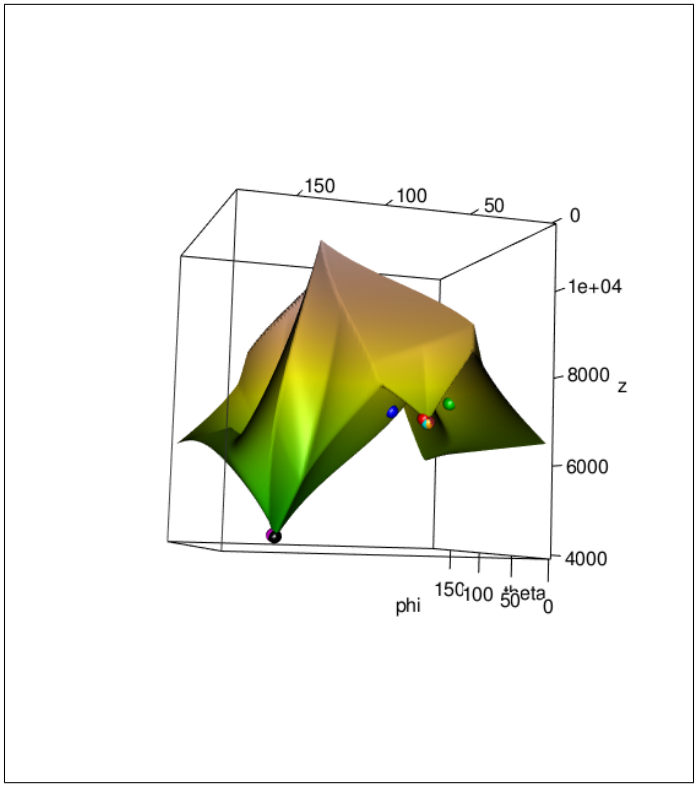


Figure 1: Plot of the objective function for the L1-norm best-fit line problem in expression (1) for a problem with 110 points in three dimensions as a function of the angles θ and ϕ that the line makes with the positive x_1 axis in the $x_1 - x_2$ plane and positive x_3 axis in the $x_1 - x_3$ plane. The plot marks the solutions found by traditional PCA (red), L1-PCA [16, 17] (cyan), L1-PCAhp [24] (orange), PCA-L1 [10] (blue), wPCA [25] (pink), awPCA [25] (brown), projection pursuit PCA (pcaPP) [27] (green), L1-PCA* [20] (magenta), and the line-fitting method `SharpEl` proposed herein (black).

$(\lambda_{ij}^{+*}, \lambda_{ij}^{-*})$ reflecting the distance along unit direction e_j between the point and its projection. If $\lambda_{ij}^{+*} + \lambda_{ij}^{-*} > 0$, we say that observation i *uses* unit direction e_j to project onto the line defined by v^*.

3. Projecting Points onto Lines

A fundamental concept in our procedure to estimate best-fit lines for point sets is the L1-norm projection. In this section we present results that will be used in the contributions in this work about projecting individual points in $\Re^m$ on specified lines through the origin: $\lambda_{ij}^+ + \lambda_{ij}^- = |x_{ij}|$ for all i, j. Since we focus on only one observation or point, we drop the index i in this section.

Proposition 2. *Let $v \neq 0$ be a given vector in $\Re^m$ such that $v_{\hat{\jmath}} \neq 0$; $\hat{\jmath} \in \{1, \ldots, m\}$. Then there is a path from the point $x \in \Re^m$ to the line generated by v using at most $m-1$ unit directions e_j; $j \neq \hat{\jmath}$.*

Proof. A proof is in the Appendix. □

In the special case when both $v_{\hat{\jmath}} = 0$ and $x_{\hat{\jmath}} = 0$ then $\lambda_j^+ = \lambda_j^- = 0$. When $v_{\hat{\jmath}} = 0$ and $x_{\hat{\jmath}} \neq 0$ then for any projection on the line defined by v, the path from the point to the projection will necessarily use the unit direction $e_{\hat{\jmath}}$, and $\lambda_{\hat{\jmath}}^+ + \lambda_{\hat{\jmath}}^- = |x_{\hat{\jmath}}|$.

According to Proposition 2, there is a point on the line defined by v which can be reached from any point using a specified subset of $m-1$ unit directions as long as $v_{\hat{\jmath}} \neq 0$ where $\hat{\jmath}$ is the unused direction. This location on the line is found by solving a system of equations. The next result establishes that one of the m subsets of $m-1$ unit directions will produce an L1-norm projection of a point on the line defined by v.

Proposition 3. *Let $v \neq 0$ be a given vector in $\Re^m$. There is an L1-norm projection of the point $x \in \Re^m$ on the line generated by v that can be reached by using at most $m-1$ unit directions. Moreover, if the unused unit direction is $e_{\hat{\jmath}}$ and $x_{\hat{\jmath}} \neq 0$, then $v_{\hat{\jmath}} \neq 0$.*

Proof. A proof is in the Appendix. □

An L1-norm projection that has an unused unit direction $e_{\hat{\jmath}}$ is said to *preserve* the $\hat{\jmath}$-th coordinate of the point x.

An essential realization about L1-norm projections of points in $\Re^m$ onto lines is that different points do not necessarily use the same subset of $m-1$ unit directions to project onto a given line. This is easily illustrated with a small example in three-dimensions. Take the line defined by $v = (1,1,1)^T$ and the three points $(2,1,3)^T$, $(3,2,1)^T$, $(1,3,2)^T$. There are projections such that the points use directions $\{e_2, e_3\}$, $\{e_1, e_3\}$, and $\{e_1, e_2\}$, and no projections where all points use the same set of two unit directions. This is in stark contrast with what occurs when projecting onto a hyperplane where there is an L1-norm projection of any point using (at most) a single unit direction and this direction is the same for any point in $\Re^m$.

4. Estimating an L1-norm best-fit line

An estimate of an L1-norm best-fit line results from modifying the nonlinear program in (2) which, as has been established, is equivalent to (1). The modification is to impose the preservation of one of the coordinates, $\hat{\jmath}$, in the projections of the n data points which means each point will use the same $m-1$ unit directions to project onto the line defined by v. As we will see, this condition transforms (2) into a linear program.

The basic step of the procedure requires fixing one of the m dimensions which will be preserved in the projections. Let this be $\hat{\jmath}$-th dimension. By Proposition 3, if $x_{i\hat{\jmath}} \neq 0$ for any i, then $v_{\hat{\jmath}} \neq 0$. Therefore, without restricting the line that will be defined by the vector v, we can set $v_{\hat{\jmath}} = 1$. This amounts to a simple normalization of the vector v; the rest of its components remain variable. From this we get $\alpha_i = x_{i\hat{\jmath}}$ for $i = 1, \ldots, n$ in the constraints in (2) and the formulation becomes:

$$z_{\hat{\jmath}} = \min_{\substack{v \in \Re^m, v_{\hat{\jmath}}=1, \\ \lambda^+, \lambda^- \in (\Re^{m \times n})^+}} \sum_{i=1}^{n} \sum_{j=1}^{m} (\lambda_{ij}^+ + \lambda_{ij}^-); \tag{3}$$

subject to:

$$v_j x_{i\hat{\jmath}} + \lambda_{ij}^+ - \lambda_{ij}^- \quad = \quad x_{ij}, \quad i = 1, \ldots, n, \; j = 1, \ldots, m; \; j \neq \hat{\jmath};$$

which is an LP. Each of the n data points generates $m-1$ constraints in this LP. An optimal solution defines a vector v such that the sum of the L1-norm distances of paths from the points to points on the line using the $m-1$ directions that preserve the $\hat{\jmath}$-th component is minimized.

Each of the m dimensions defines a different LP; one for each subset of $m-1$ unit directions that can be used exclusively to project the data. The procedure we propose to estimate an L1-norm best-fit line to the data is based on solving these m LPs and selecting the vector v from the solution associated with the smallest of the m objective function values.

Note that if $x_{i\hat{\jmath}} = 0$ for all points i, then the points reside in an $(m-1)$-dimensional subspace. The LP has an optimal solution that corresponds to projecting all points to the origin. The objective function value is an upper bound for the optimal objective function values for any of the LPs, as it has the same objective function value as the feasible solution $v = 0$.

Procedure `SharpEl` for estimating an L1-norm best-fit line.
Given points $x_i \in \Re^m$ for $i = 1, \ldots, n$.

1: Set $z^* = j^* = \infty$,.
2: **for** $(\hat{\jmath} = 1;\ \hat{\jmath} \leq m;\ \hat{\jmath} = \hat{\jmath} + 1)$ **do**
3: Solve LP in (3).
4: **if** $z_{\hat{\jmath}} < z^*$, **then**
5: Set $z^* = z_{\hat{\jmath}};\ j^* = \hat{\jmath}$.
6: **end if**
7: **end for**

Procedure `SharpEl` identifies the component, $\hat{\jmath} = j^*$, which defines an LP in (3). This LP has the smallest objective function value from among the m possible LPs that result from the data as the preserved dimension ranges from 1 to m. The vector v in the optimal solution of this LP defines the estimated line of which its $\hat{\jmath}$-th component is 1 and on which the i-th point's projection occurs at $vx_{i\hat{\jmath}}$.

Consider again the three points $(2,1,3)^T$, $(3,2,1)^T$, $(1,3,2)^T$. Procedure `SharpEl` generates a solution $v = (3/2, 1, 9/4)^T$ with objective function value

$z = 20/3$. A global optimal solution is $v^* = (1, \sqrt{3/2}, 3/2)^T$ with objective function value $z^* = (13 + 2\sqrt{6})/3$.

Procedure `SharpEl` can attain globally optimal solutions for certain instances. Figure 2 contains a plot of the points that were used to generate Figure 1. Procedure `SharpEl` generates a solution $v^* = (0.091, 0.704, -0.705)^T$ with objective function $z^* = 4192.3$. The solution corresponds to the globally optimal solution at $\theta = 82.6°$ and $\phi = 134.8°$ in Figure 1.

Procedure `SharpEl` requires finding the solution to m LPs each with $n(m-1)$ constraints. Although solving LPs is efficient, the LPs formulated here can be large and solving them directly using an LP solver can be computationally demanding and time consuming. Remarkably, it is possible to solve the LPs in (3) by sorting simple ratios.

Proposition 4. *An optimal solution v^* to the LP in* (3) *can be constructed as follows. If $x_{i\hat{\jmath}} = 0$ for all i, then set $v^* = 0$. Otherwise, for each $j \neq \hat{\jmath}$,*

1. *Take points i such that $x_{i\hat{\jmath}} \neq 0$ and sort the ratios $\frac{x_{ij}}{x_{i\hat{\jmath}}}$ in increasing order.*
2. *Let $\tilde{\imath}$ be the index of the given point such that*

$$\sum_{i:i>\tilde{\imath}} |x_{i\hat{\jmath}}| < \frac{1}{2} \sum_{i=1}^{n} |x_{i\hat{\jmath}}|,$$

and

$$\sum_{i:i<\tilde{\imath}} |x_{i\hat{\jmath}}| \leq \frac{1}{2} \sum_{i=1}^{n} |x_{i\hat{\jmath}}|.$$

3. *Set $v_j^* = \frac{x_{\tilde{\imath}j}}{x_{\tilde{\imath}\hat{\jmath}}}$.*

Proof. A proof is in the Appendix. □

Access to a solution to LP (3) provided by Proposition 4 means the instruction in Step 3 in procedure `SharpEl` requires sorting $m(m-1)$ lists of n numbers each. Sorting n numbers using standard algorithms requires $O(n \log n)$ steps, so the time complexity for procedure `SharpEl` is $O(m^2 n \log n)$. Sorting each list is independent and therefore both fixing coordinates and calculating components of v can be fully parallelized into subprocesses each requiring $O(n \log n)$ time for increased computational efficiency.

5. Extensions to Subspace Fitting and Principal Component Analysis

We can extend our estimates of an L1-norm best-fit line to derive estimates of L1-norm best-fit subspaces of other dimensions. An L1-norm best-fit subspace of dimension q is an optimal solution to the following problem:

$$\min_{\substack{V \in \Re^{m \times q}, \alpha_i \in \Re^q; \\ i=1,\ldots,n}} \sum_{i=1}^{n} \|x_i - V\alpha_i\|_1. \tag{4}$$

The singular value decomposition of the data matrix X, whose rows are the x_i, provide an optimal solution for the sum of squares analogue to this problem. The singular value decomposition for a column-wise mean-centered data matrix produces principal components for traditional PCA.

Our method can be extended to estimate a q-dimensional subspace by iteratively fitting one-dimensional subspaces and projecting data into orthogonal subspaces. Let $X = X^1$ and let V^k be the matrix whose columns are the best-fit one-dimensional subspaces. The matrix V^1 will have one column corresponding to the estimate of the best-fit line. For $k = 2, \ldots, q$, project data into a subspace orthogonal to the fitted subspaces:

$$X^k = X^{k-1} - X^{k-1}V^{k-1}(V^{k-1})^T.$$

Then estimate the best-fit line for X^k and append it to the matrix V^{k-1} to form V^k. The estimate for the best-fit line may not reside in the subspace containing the data, and so we need to project out the directions already covered by V^{k-1}:

$$V^k = V^k - V^kV^{k-1}(V^{k-1})^T.$$

In the computational experiments, we will refer to this procedure as `SharpEl1-PCA`.

6. Computational Results and Analysis

By Proposition 1, each point projects onto a line using $m-1$ unit directions, but not all points use the same $m-1$ unit directions. The number of possible combinations of directions is $\binom{n}{m-1} = O(n^{m-1})$. In procedure `SharpEl`

we force all points to use the same $m-1$ unit directions. This generates just m possibilities. In this section we demonstrate that checking this small subset of unit directions produces high-quality solutions. The lines produced by procedure `SharpEl` ignore the effect of outliers better than traditional PCA and previously-proposed L1-norm line-fitting heuristics.

6.1. Robust Line-Fitting in Synthetic Data

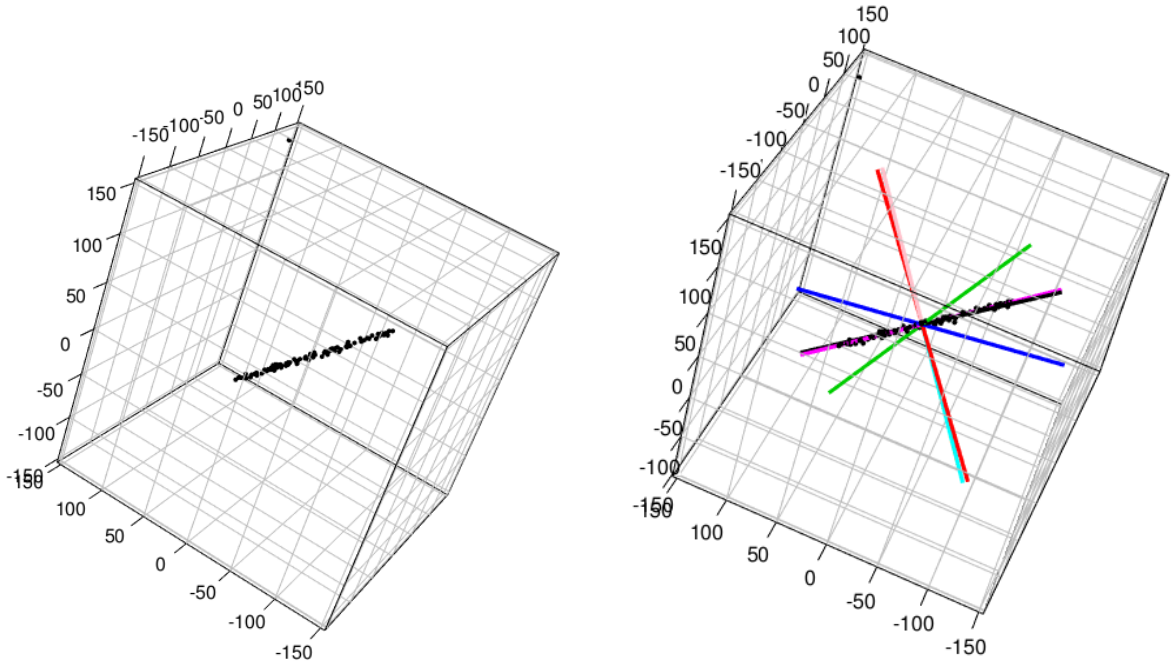
In this section we compare the ability of procedure `SharpEl` for estimating an L1-norm best-fit line to traditional PCA (PCA) and other methods based on the L1 norm using synthetic data. For each of ten replications, we create datasets with $n = 100$ observations including 10 outliers. For each replication, each coordinate of the "true" v is sampled from a Uniform $(-1, 1)$ distribution and after all coordinates have been sampled, v is normalized so that $\|v\|_2 = 1$. The experiment is repeated for dimensions $m = 3$, 10, 100, 500.

For non-outlier observations, each "true" α is sampled from a Uniform $(-100, 100)$ distribution to locate the projection on the line. Noise following a Laplace(0,1) distribution is added to locate the points off of the line. Outlier observations are created by sampling the first five coordinates from a Uniform $(100, 150)$ distribution and noise sampled from a Laplace(0,1) distribution is added.

Our experiment is conducted in the R Environment for Statistical Computing [28]. R code and data are included as supplementary files. We compare our method to traditional PCA in the function *prcomp()*. We also compare our method to eight L1-norm PCA methods. One is implemented in the R package pcaPP [29, 27] and is a method based on projection pursuit. The remaining methods are implemented in the R package pcaL1 [30]. L1-PCA [16, 17] is based on alternate optimization, fixing v and solving for α and fixing α and solving for v in (1). L1-PCAhp [24] employs a heuristic for estimating an L1-norm best-fit hyperplane at each iteration. PCA-L1 [10] is a method to estimate a line that maximizes the sum of the L1-norm lengths of the projections. wPCA and awPCA [25] are methods that use iteratively re-weighted least squares al-

gorithms to minimize the sum of L1-norm distances to L2-norm projections in a subspace. L1-PCA* [20] is a method based on iteratively finding L1-norm best-fit hyperplanes.

The performance of each method is evaluated by measuring discordance between the vector returned by the method and the vector defined by the true line. Discordance is measured as $1 - v^T v'$, where v defines the true line and v' is the vector found by a particular method.



(a) (b)

Figure 2: (a) Example of synthetic data set for $m = 3$ with points sampled along a line with noise plus a set of clustered leverage points as outliers, and (b) the best-fit lines given by traditional PCA (red), L1-PCA [16, 17] (cyan), L1-PCAhp [24] (orange), PCA-L1 [10] (blue), wPCA [25] (pink), awPCA [25] (brown), projection pursuit PCA (pcaPP) [27] (green), L1-PCA* [20] (magenta), and the `SharpEl` heuristic (black).

Figure 2(a) depicts one of the datasets in three dimensions. Most of the points lie near the line and the rest are clustered leverage outliers in the top left of the plot. Figure 2(b) depicts the lines found by PCA, L1-PCA, L1-PCAhp,

PCA-L1, wPCA, awPCA, pcaPP, L1-PCA*, and procedure `SharpEl`. PCA, L1-PCA, L1-PCAhp, wPCA, and awPCA are dramatically affected by the outlier observations. PCA-L1 and pcaPP produce a line that clearly reflects error due to the outlier observations, though the effect is not as severe as for PCA. L1-PCA* and procedure `SharpEl` appears to ignore the outlier observations and fits the non-outlier observations and is a good representation of the "true" line.

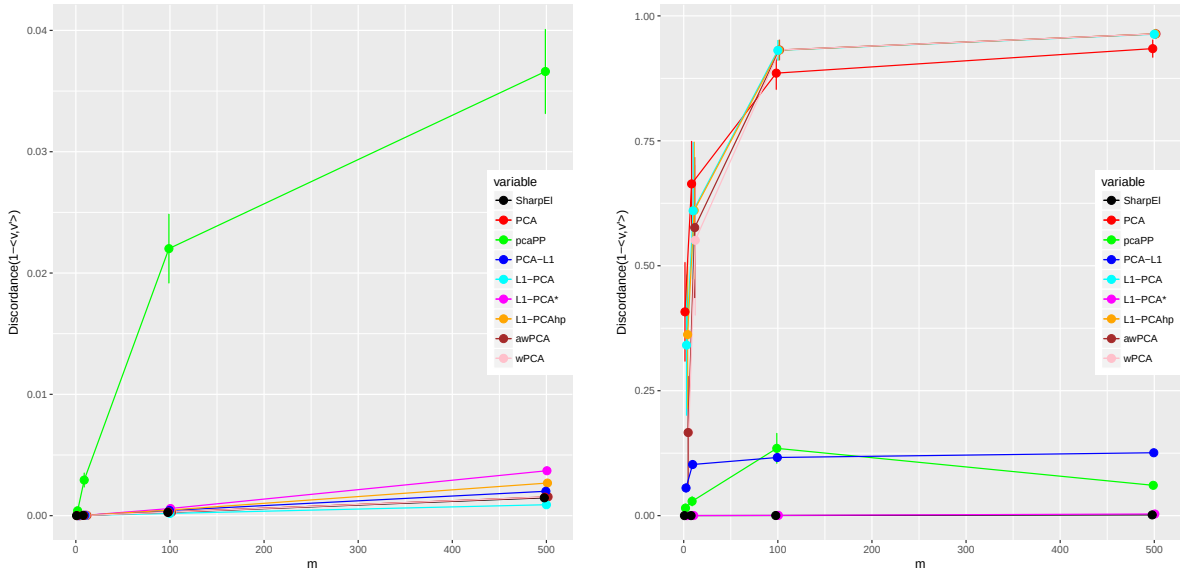


Figure 3: Mean discordance between the true line v and the fitted line v', measured as $1-v^T v'$, for different dimensions of the data m (a) without and (b) with outlier contamination for PCA, L1-PCA [16, 17], L1-PCAhp [24], PCA-L1 [10], wPCA [25], awPCA [25], pcaPP [29], L1-PCA* [20], and the `SharpEl` heuristic. Error bars reflect the standard errors for each combination of method and value of m.

Figure 3 contains plots of the mean discordance as a function of the number of dimensions m for each dataset (a) without and (b) with outlier-contamination. In the absence of outlier contamination, all of the methods have low discordance as indicated by the scale. Procedure pcaPP has larger values than the other methods.

With outlier contamination, `SharpEl` and L1-PCA* find better lines than the

other methods for every value of m. The lines estimated by these two methods have nearly zero discordance with the true line. PCA, L1-PCA, L1-PCAhp, PCA-L1, wPCA, and awPCA produce lines with errors that increase with m until $m > n$ at which point the discordance plateaus. The errors for pcaPP decrease after $m > n$ and as m increases. For small m, PCA, L1-PCA, L1-PCAhp, wPCA, and awPCA exhibit large standard errors in the discordance, reflecting a sensitivity to the particular true line and/or sample of points. For $m > n$, these methods appear to fit the outlier observations and produce lines that are nearly orthogonal to the true lines. PCA-L1 and pcaPP exhibit similar behavior, but never exceed a discordance of 0.15.

We attribute the superior performance of `SharpEl` and L1-PCA* to the fact that these methods do not rely on an initial starting point before converging to a local optimal solution. Rather, they both exploit particular problem structures to obtain good estimates of the optimal objective function values. In contrast, the other L1-norm based methods require an initial starting point. The objective functions for the associated L1-norm optimization problems have large numbers of local minima, and these methods appear to be converging to poor local minima.

6.2. Forecasting Call Center Arrivals

We apply our method and competing methods to data collected for a call center and reported by Shen and Huang [31]. The goal is to forecast future call volume based on historical data. The dataset includes call volume for 200 days of data, and 68 15-minute time periods during each day. A forecasting method proposed by Shen and Huang [31] is to find the singular value decomposition of historical call data, use the first four basis vectors to represent the data, fit time-series models for each vector of scores to predict multipliers for the loadings for the next day, and apply the multipliers to the loadings to generate a forecast. Forecasts are generated based on rolling windows of 150 days for days 151 to 200 and compared to the actual recorded call volume. Recall the authors suppress outliers in their data prior to their analysis [31].

In our analysis, we used the data from [31] without and with outlier contamination. To introduce outlier contamination, we sampled days where the call volume for the last 15 time periods of each day was higher than usual. Let $\bar{x} \in \Re^{68}$ be the sample mean and let S be the sample covariance matrix for the original data. We sampled 10 outlier days, one for each day of two five-day work weeks, from a multivariate normal distribution with mean μ and covariance matrix S where,

$$\begin{aligned} \mu_i &= \bar{x}_i, i = 1, 2, \ldots, 53, \\ \mu_{54+i} &= 2\bar{x}_{12+i(\text{mod } 5)}, i = 0, \ldots, 14. \end{aligned}$$

The mean call volume for the last 15 time periods of each day are replaced by twice the mean during the peak times (time periods 12 to 16). Outliers therefore comprised 4.8% of the complete dataset.

The increased call volume during later periods in the day in the outlier days could represent a situation where the company creates incentives to boost activity during typically slow periods. Long term, it is in the manager's interest to staff the call center with the typical call volume rather than the increased activity in the later periods. Staffing according to the increased call volume will result in excessive idle times for employees, and additional costs for additional works. To prepare for special events such as incentives, part-time staff could be employed.

Singular value decomposition (PCA without centering data) is used to generate scores on the first four principal components as described for the TS4 method in [31]. For the L1-norm PCA methods, scores are generated by applying the PCA method to uncentered data to find basis vectors for a fitted four-dimensional subspace. L1-norm projections of points on to the subspace are determined by solving an LP by fixing v in (2) and solving for the α_i, $i = 1, \ldots, n$. These numbers are used as the scores. For each of the four score vectors, an AR(1) model is fit with a day-of-the-week effect [31] and used to generate a prediction for a multiplier for the loadings for the forecasted day. As in the original work, we apply the root-unroot method [31]. The mean relative

Table 1: Summary Statistics for RMSE and MRE for Forecasts with No Outliers

	RMSE				MRE			
	1st Qu.	Median	Mean	3rd Qu.	1st Qu.	Median	Mean	3rd Qu.
PCA	**42.7**	54.0	61.9	69.7	5.4	6.6	**7.4**	8.5
L1-PCA [16, 17]	42.9	54.0	61.9	67.9	5.3	6.6	**7.4**	8.5
L1-PCAhp [24]	42.9	54.2	61.7	68.3	**5.0**	**6.5**	7.5	9.0
PCA-L1 [10]	**42.7**	53.0	61.8	68.5	5.4	6.6	**7.4**	**8.4**
wPCA [25]	80.4	91.4	94.3	101.3	8.8	9.6	10.4	11.2
awPCA [25]	130.3	166.2	152.9	185.6	16.3	25.6	22.6	28.4
pcaPP [27]	590.7	645.1	661.9	717.1	67.8	72.5	72.1	75.5
L1-PCA* [20]	**42.7**	53.4	61.6	**67.4**	**5.0**	6.6	7.5	8.8
`SharpEl1-PCA`	43.6	**52.3**	**61.4**	67.5	5.2	6.6	**7.4**	8.6

error (MRE) and root mean squared error (RMSE), measuring the difference between the forecasted call volume and the actual call volume, are recorded for each forecast.

Summary statistics for MRE and RMSE are included in Tables 1 and 2. For the data without outliers, there is little difference among the proposed `SharpEl1-PCA` and PCA, L1-PCA, L1-PCAhp, PCA-L1, and L1-PCA*. The median RMSE was slightly better for `SharpEl1-PCA`, PCA-L1, and L1-PCA* than PCA. For `SharpEl1-PCA`, L1-PCA, L1-PCAhp, and L1-PCA*, the median MRE was within 0.1% of that for PCA. The errors for wPCA and awPCA were noticably higher than the other methods; the median MRE values were 1.5-3.9 times as large as for PCA and the median RMSE values were 1.7-3.1 times as large as for PCA. The errors for pcaPP were dramatically higher than for PCA; the median MRE was 11.0 times as large and the RMSE was 11.9 times as large as for PCA.

When the outlier contamination is added, the median error rates actually decrease for `SharpEl1-PCA`, L1-PCA, and L1-PCAhp. Because they are effectively ignoring the outliers, the largest errors are for the few outlier observations

Table 2: Summary Statistics for RMSE and MRE for Forecasts with Outliers

	RMSE				MRE			
	1st Qu.	Median	Mean	3rd Qu.	1st Qu.	Median	Mean	3rd Qu.
PCA	55.5	65.0	75.8	83.8	12.5	13.8	16.0	19.1
L1-PCA [16, 17]	55.4	64.3	75.6	84.0	12.4	13.9	16.0	19.1
L1-PCAhp [24]	43.6	51.1	62.6	**69.8**	**4.9**	**6.2**	7.6	9.6
PCA-L1 [10]	50.9	59.5	71.5	82.0	10.7	12.2	14.3	17.2
wPCA [25]	1011.0	1061.0	1078.0	1125.0	96.0	185.7	165.8	207.7
awPCA [25]	994.8	1017.0	1031.0	1103.0	89.9	91.4	90.7	93.2
pcaPP [27]	561.2	599.2	607.0	655.1	69.8	73.8	73.7	77.3
L1-PCA* [20]	43.6	50.1	62.4	70.5	**4.9**	**6.2**	7.6	9.6
`SharpEl1-PCA`	**43.3**	**49.3**	**61.9**	70.0	5.0	6.3	**7.5**	**9.2**

that comprise only 4.8% of the data. The median MRE values for these methods are reductions of 54.3-55.1% from the median MRE for PCA and the median RMSE values are reductions of 21.3-24.1% from those for PCA. PCA-L1 had error rates slightly better than PCA and L1-PCA performed roughly the same as PCA, indicating influence by the outliers. The performances of wPCA and awPCA degraded in the presence of outliers with median MRE values 13.4 and 6.6 times that of PCA and median RMSE values 16.3 and 15.6 times that of PCA. The performance of pcaPP was slightly better in the presence of outlier contamination, but still worse than all methods except for wPCA and awPCA.

The poor performance of L1-PCA, PCA-L1, wPCA, and awPCA may be attributed to the fact that they are locally-convergent methods that require an initial starting point guess. In our implementations, they use PCA to derive a starting point, and appear to converge to local solutions with similar performance.

7. Conclusions

In recent literature, the problems arising from outliers in data when using methods based on minimizing the sum of squared errors is acknowledged as a serious limitation to these methods. The response has been an increase in interest in robust methods and hence using the L1 norm to replace the sum of squares criterion in many of the standard methods. Using the L1 norm allows outliers to be part of the data without unduly affecting the final results.

Many analytics methods, including linear regression, logistic regression, and traditional PCA, require fitting subspaces to (centered) data to extract information about properties such as location, dispersion, and orientation. A fitted subspace simplifies analyses of data including predictions. Except for the case of a hyperplane and a point, finding an L1-norm best-fit subspace for a point set in m-dimensions is a nonlinear, nonconvex, nonsmooth optimization problem.

We introduce an efficient procedure, `SharpEl`, to estimate a best-fit line using only LPs. An easy way to construct solutions to the LPs based on sorting ratios also makes `SharpEl` fast and scalable. Procedure `SharpEl` produces estimates of the L1-norm best-fit line to data that perform substantially better than the other competing approaches in numerical tests. In fact, at times, the estimates are actually optimal. The estimated lines generated with `SharpEl` are the basis for the "forward" L1-norm PCA procedure `SharpEl1-PCA`.

Procedure `SharpEl1-PCA` was applied to data which had been previously used in a study to forecast call center arrivals. Computational comparisons using the data with demand surges (outliers) with competing methods show that our robust method performs well. Also, we observed that if data do not contain outliers, then the performance of robust methods do not suffer. Based on these considerations, if there is a risk that data contain outlier observations, analytics models should be built using robust techniques such as the one proposed here.

References

[1] I. T. Jolliffe, Principal Component Analysis, 2nd Edition, Springer, 2002.

[2] J. P. Brooks, R. A. Reris, Principal Component Analysis and Optimization: A Tutorial, Operations Research & Computing: Algorithms & Software for Analytics, 2015, pp. 212–225.

[3] H. Martini, A. Schöbel, Median hyperplanes in normed spaces - a survey, Discrete Applied Mathematics 89 (1998) 181–195.

[4] J. P. Brooks, J. H. Dulá, The L1-norm best-fit hyperplane problem, Applied Mathematics Letters 26 (2013) 51–55.

[5] V. Boltyanski, H. Martini, V. Soltan, Median Problems in Location Science, Geometric Methods and Optimization Problems, Springer, 1999, pp. 231–355.

[6] G. Bruno, A. Genovese, G. Improta, A historical perspective on location problems, BSHM Bulletin: Journal of the British Society for the History of Mathematics 29 (2) (2014) 83–97.

[7] N. Megiddo, A. Tamir, Finding least-distance lines, SIAM Journal of Algebraic Discrete Methods 4 (1983) 207–211.

[8] J. Brimberg, H. Juel, A. Schöbel, Properties of three-dimensional median line location models, Annals of Operations Research 122 (2003) 71–85.

[9] R. Blanquero, E. Carrizosa, A. Schöbel, D. Scholz, A global optimziation procedure for the location of a median line in the three-dimensional space, European Journal of Operational Research 215 (2011) 14–20.

[10] N. Kwak, Principal component analysis based on L1-norm maximization, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 1672–1680.

[11] F. Nie, H. Huang, C. Ding, D. Luo, H. Wang, Robust principal component analysis with non-greedy ℓ_1-norm maximization, Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (2011) 1433–1438.

[12] S. Kundu, P. P. Markopoulos, D. A. Pados, Fast computation of the L1-principal component of real-valued data, IEEE International Conference on Acoustic, Speech, and Signal Processing (2014) 8028–8032.

[13] P. P. Markopoulos, K. N. Karystinos, D. A. Pados, Optimal algorithms for L_1-subspace signal processing, IEEE Transactions on Signal Processing 62 (2014) 5046–5058.

[14] E. J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis?, Journal of the ACM 58 (3) (2011) 11:1–11:37.

[15] D. Goldfarb, S. Ma, K. Scheinberg, Fast alternating linearization methods for minimizing the sum of two convex functions, Mathematical Programming 141 (2013) 349–382.

[16] Q. Ke, T. Kanade, Robust subspace computation using L1 norm, Tech. rep., Carnegie Mellon University (2003).

[17] Q. Ke, T. Kanade, Robust L1 norm factorization in the presence of outliers and missing data by alternative convex programming, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2005.

[18] N. Tsagkarakis, P. P. Markopoulos, D. A. Pados, On the l_1-norm approximation of a matrix by another of lower rank, 15th IEEE International Conference on Machine Learning and Applications (ICMLA) (2016) 768–773.

[19] F. Jiang, O. Enqvist, F. Kahl, A combinatorial approach to L1-matrix factorization, Journal of Mathematical Imaging and Vision 51 (2015) 430–441.

[20] J. P. Brooks, J. H. Dulá, E. L. Boone, A pure L1-norm principal component analysis, Computational Statistics & Data Analysis 61 (2013) 83–98.

[21] J. Damon, J. S. Marron, Backwards principal component analysis and principal nested relations, Journal of Mathematical Imaging and Vision 50 (1-2) (2014) 107–114.
URL http://dx.doi.org/10.1007/s10851-013-0463-2

[22] K. L. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Mahoney, X. Meng, D. P. Woodruff, The fast Cauchy transform and faster robust linear regression, Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms (2013) 466–477.

[23] K. L. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Mahoney, X. Meng, D. P. Woodruff, The fast Cauchy transform and faster robust linear regression, SIAM Journal on Computing 45 (3) (2016) 763–810.

[24] A. Visentin, S. Prestwich, S. A. Tarim, Robust principal component analysis by reverse iterative linear programming, Joint European Conference on Machine Learning and Knowledge Discovery in Databases (2016) 593–605.

[25] Y. W. Park, D. Klabjan, Iteratively reweighted least squares algorithms for L1-norm principal component analysis, IEEE International Conference on Data Mining (ICDM)Journal: Submitted.

[26] N. Gillis, S. A. Vavasis, On the complexity of robust PCA and L1-norm low-rank matrix approximation, arXiv 1509.09236v2 [cs.LG].

[27] C. Croux, A. Ruiz-Gazen, High breakdown estimators for prinicpal components: The projection-pursuit approach revisited, Journal of Multivariate Analysis 95 (2005) 206–226.

[28] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2016).
URL https://www.R-project.org/

[29] P. Filzmoser, H. Fritz, K. Kalcher, pcaPP: Robust PCA by projection pursuit, 2009.
URL http://cran.r-project.org/web/packages/pcaPP/index.html

[30] S. Jot, J. P. Brooks, A. Visentin, Y. W. Park, pcal1: L1-norm PCA methods.
URL https://cran.r-project.org/web/packages/pcaL1/index.html;

[31] H. Shen, J. Z. Huang, Interday forecasting and intraday updating of call center arrivals, Manufacturing & Service Operations Management 10 (3) (2008) 391–410.

[32] A. Charnes, W. W. Cooper, R. O. Ferguson, Optimal estimation of executive compensation by linear programming, Management Science 1 (1955) 138–150.

Appendix. Proofs of Results.

Proposition. *1. Formulation* (2) *is equivalent to* (1).

Proof. The optimization problem in (2) follows from applying a series of substitutions and algebraic manipulations that maintain the equivalence with (1) at every step. To deal with the sum of absolute values of functions in the objective function we apply a technique originally proposed by Charnes and Cooper [32]. It consists of substituting each term of the summation in the objective function by a pair of nonnegative "goal" variables: $x_{ij} - v\alpha_i = \lambda_{ij}^+ - \lambda_{ij}^-,\ \lambda_{ij}^+, \lambda_{ij}^- \geq 0;\ \forall i\&j$; these become the constraints. The new objective function, $\sum_{i=1}^{n}\sum_{j=1}^{m}|\lambda_{ij}^+ - \lambda_{ij}^-|$ can be replaced with $\sum_{i=1}^{n}\sum_{j=1}^{m}(\lambda_{ij}^+ + \lambda_{ij}^-)$ due to the fact that only one in all pairs $\lambda_{ij}^+, \lambda_{ij}^-, \forall i\&j$ can be positive for some optimal solution. Any feasible solution for (2) generates an objective function value which is the same as that of (1) using the same values for v and α; and vice versa. Therefore, an optimal solution to (1) generates a feasible solution for (2) which would have to be optimal since otherwise there would be a contradiction; and vice-versa. □

Proposition. *2. Let $v \neq 0$ be a given vector in $\Re^m$ such that $v_{\hat{\jmath}} \neq 0$; $\hat{\jmath} \in \{1, \ldots, m\}$. Then there is a path from the point $x \in \Re^m$ to the line generated by v using at most $m-1$ unit directions e_j; $j \neq \hat{\jmath}$.*

Proof. For a given $v \neq 0$ the expression in (2) is a linear program. By setting $\alpha = x_{\hat{\jmath}}/v_{\hat{\jmath}}$ in (2), we can generate values for the remaining variables and set $\lambda^+_{\hat{\jmath}} = \lambda^-_{\hat{\jmath}} = 0$. Notice that this corresponds to a basic feasible solution. □

Proposition. *3. There is an L1-norm projection of the point $x \in \Re^m$ on the line generated by v that can be reached by using at most $m-1$ unit directions. Moreover, if the unused unit direction is $e_{\hat{\jmath}}$ and $x_{\hat{\jmath}} \neq 0$, then $v_{\hat{\jmath}} \neq 0$.*

Proof. Given v, the projection of a point x is found by solving a linear program corresponding to a subset of the constraints of (2). Any basic feasible solution will have α as basic because it is unrestricted in sign. The remaining $m-1$ basic variables will be one from each pair $(\lambda^+_j, \lambda^-_j)$ for $j \neq \hat{\jmath}$ for some $\hat{\jmath}$. The determinant of the basis is $\pm v_{\hat{\jmath}}$ and therefore this basis is non-singular if and only if $v_{\hat{\jmath}} \neq 0$. □

Proposition. *4. An optimal solution v^* to the LP in (3) can be constructed as follows. For each $j \neq \hat{\jmath}$,*

1. *Take points i such that $x_{i\hat{\jmath}} \neq 0$ and sort the ratios $\frac{x_{ij}}{x_{i\hat{\jmath}}}$ in increasing order.*
2. *Let $\tilde{\imath}$ be the index of the given point such that*

$$\sum_{i:i>\tilde{\imath}} |x_{i\hat{\jmath}}| < \frac{1}{2} \sum_{i=1}^{n} |x_{i\hat{\jmath}}|,$$

and

$$\sum_{i:i<\tilde{\imath}} |x_{i\hat{\jmath}}| \leq \frac{1}{2} \sum_{i=1}^{n} |x_{i\hat{\jmath}}|.$$

3. *Set $v^*_j = \frac{x_{\tilde{\imath}j}}{x_{\tilde{\imath}\hat{\jmath}}}$.*

Proof. First, note that if $x_{i\hat{\jmath}} = 0$ for a point i, then the constraints for point i in (3) are of the form

$$\lambda^+_{ij} - \lambda^-_{ij} = x_{ij}, j \neq \hat{\jmath},$$

and the contribution to the objective function value for that point will be a constant value equal to $\sum_{j \neq \hat{\jmath}} |x_{ij}|$ for any v, so we may exclude them from consideration in deriving v^*. If $x_{i\hat{\jmath}} = 0$ for all points i, then the solution $v = 0$ achieves an objective function value of $\sum_{i=1}^{n} \sum_{j \neq \hat{\jmath}} |x_{ij}|$ and is therefore optimal.

Similarly, if $x_{i\hat{\jmath}} = 0$ for a particular point i, then the contribution to the objective function is $\sum_{j=1}^{m} |x_{ij}|$ regardless of the value for v. Therefore, the optimization of (3) depends only on choosing v based on points i for which $x_{i\hat{\jmath}} \neq 0$ and we assume that this is true for all points hereafter.

For a given j, let $\tilde{x}_j = x_{\tilde{\imath}j}$. Note that we can rewrite (1) as

$$\min_{\substack{v \in \Re^m, \alpha_i \in \Re; \\ i=1,\ldots,n}} \sum_{i=1}^{n} \sum_{j=1}^{m} |x_{i\hat{\jmath}}| \left| \frac{x_{ij}}{x_{i\hat{\jmath}}} - v \right|,$$

and therefore (3) can be written as

$$z_{\hat{\jmath}} = \min_{\substack{v \in \Re^m, v_{\hat{\jmath}} = 1, \\ \lambda^+, \lambda^- \in (\Re^{m \times n})^+}} \sum_{i=1}^{n} \sum_{j=1}^{m} |x_{i\hat{\jmath}}| (\lambda_{ij}^+ + \lambda_{ij}^-); \tag{5}$$

subject to:

$$v_j + \lambda_{ij}^+ - \lambda_{ij}^- \quad = \quad \tfrac{x_{ij}}{x_{i\hat{\jmath}}}, \quad i = 1, \ldots, n, \; j = 1, \ldots, m; \; j \neq \hat{\jmath};$$

For a given point i and attribute $j \neq \hat{\jmath}$, because the coefficients for λ_{ij}^+ and λ_{ij}^- are linearly dependent in (5), at most one will be nonzero in a basic feasible solution. Therefore

$$\lambda_{ij}^+ + \lambda_{ij}^- = \left| \frac{x_{ij}}{x_{i\hat{\jmath}}} - \frac{\tilde{x}_j}{\tilde{x}_{\hat{\jmath}}} \right|,$$

and the primal objective function value is

$$\sum_{i=1}^{n} \sum_{j=1}^{m} |x_{i\hat{\jmath}}| \left| \frac{x_{ij}}{x_{i\hat{\jmath}}} - \frac{\tilde{x}_j}{\tilde{x}_{\hat{\jmath}}} \right|.$$

We will construct a dual feasible solution to (5) with the same objective function value as that for v^*.

The dual of (5) using this form is

$$\max \sum_{i=1}^{n} \sum_{j \neq \hat{\jmath}} \frac{x_{ij}}{x_{i\hat{\jmath}}} \pi_{ij},$$

s.t.

$$\sum_{i=1}^{n} \pi_{ij} = 0,\ j = 1, \ldots, m,$$

$$-|x_{i\hat{\jmath}}| \leq \pi_{ij} \leq |x_{i\hat{\jmath}}|,\ i = 1, \ldots, n, j = 1, \ldots, m.$$

For each $j \neq \hat{\jmath}$, suppose that the ratios have been sorted as in the statement of the proposition. Set the values for the dual variables π_{ij} as follows

$$\pi_{ij} = \begin{cases} |x_{i\hat{\jmath}}| & \text{for } i > \tilde{\imath} \\ -|x_{i\hat{\jmath}}| & \text{for } i < \tilde{\imath} \end{cases},$$

and

$$\pi_{\tilde{\imath}j} = -\sum_{i \neq \tilde{\imath}} \pi_{ij}.$$

This is a dual feasible solution because of the choice of $\tilde{\imath}$ for each j. To see that $\pi_{\tilde{\imath}j} \geq -|x_{i\hat{\jmath}}|$:

$$\begin{aligned} \pi_{\tilde{\imath}j} &= -\textstyle\sum_{i \neq \tilde{\imath}} \pi_{ij}, \\ &= -\left(-\textstyle\sum_{i<\tilde{\imath}} |x_{i\hat{\jmath}}| + \sum_{i>\tilde{\imath}} |x_{i\hat{\jmath}}|\right), \\ &= \textstyle\sum_{i=1}^{n} |x_{i\hat{\jmath}}| - |x_{\tilde{\imath}\hat{\jmath}}| - 2\sum_{i>\tilde{\imath}} |x_{i\hat{\jmath}}|, \\ &\geq -|x_{\tilde{\imath}\hat{\jmath}}|. \end{aligned}$$

The inequality follows from the choice of $\tilde{\imath}$. By a similar argument, $\pi_{\tilde{\imath}j} \leq |x_{\tilde{\imath}\hat{\jmath}}|$.

The objective function value for the dual is

$$\begin{aligned} \textstyle\sum_{j=1}^{m}\sum_{i=1}^{n} \frac{|x_{ij}|}{x_{i\hat{\jmath}}}\pi_{ij} &= \textstyle\sum_{j=1}^{m}\left(\sum_{i>\tilde{\imath}} \frac{|x_{ij}|}{x_{i\hat{\jmath}}} - \sum_{i<\tilde{\imath}} \frac{x_{ij}|}{x_{i\hat{\jmath}}} - \frac{x_{\tilde{\imath}j}}{x_{\tilde{\imath}\hat{\jmath}}} \sum_{i \neq \tilde{\imath}} \pi_{ij}\right), \\ &= \textstyle\sum_{j=1}^{m}\left(\sum_{i<\tilde{\imath}} \left(-\frac{x_{ij}}{x_{i\hat{\jmath}}}|x_{i\hat{\jmath}}| + \frac{x_{\tilde{\imath}j}}{x_{\tilde{\imath}\hat{\jmath}}}|x_{i\hat{\jmath}}|\right) + \sum_{i>\tilde{\imath}} \left(\frac{x_{ij}}{x_{i\hat{\jmath}}}|x_{i\hat{\jmath}}| - \frac{x_{\tilde{\imath}j}}{x_{\tilde{\imath}\hat{\jmath}}}|x_{i\hat{\jmath}}|\right)\right), \\ &= \textstyle\sum_{j=1}^{m}\sum_{i<\tilde{\imath}} |x_{i\hat{\jmath}}| \left(\frac{x_{\tilde{\imath}j}}{x_{\tilde{\imath}\hat{\jmath}}} - \frac{x_{ij}}{x_{i\hat{\jmath}}}\right) + \sum_{j=1}^{m}\sum_{i>\tilde{\imath}} |x_{i\hat{\jmath}}| \left(-\frac{x_{\tilde{\imath}j}}{x_{\tilde{\imath}\hat{\jmath}}} + \frac{x_{ij}}{x_{i\hat{\jmath}}}\right), \\ &= \textstyle\sum_{j=1}^{m}\sum_{i=1}^{n} |x_{i\hat{\jmath}}| \left|\frac{x_{ij}}{x_{i\hat{\jmath}}} - \frac{x_{\tilde{\imath}j}}{x_{\tilde{\imath}\hat{\jmath}}}\right|. \end{aligned}$$

We have constructed a dual feasible solution with the same objective function value as our proposed primal solution, and so the solution is optimal. □