Neural Networks 84 (2016) 28–38

Contents lists available at ScienceDirect

Neural Networks

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

Piece-wise quadratic approximations of arbitrary error functions for fast and robust machine learning

A.N. Gorban [a,*], E.M. Mirkes [a], A. Zinovyev [b]

[a] Department of Mathematics, University of Leicester, Leicester, LE1 7RH, UK
[b] Institut Curie, PSL Research University, Mines Paris Tech, Inserm, U900, F-75005, Paris, France

HIGHLIGHTS

- The quadratic error functionals demonstrate many weaknesses for complex data.
- The back side of the non-quadratic error functionals is cost for optimization.
- New algorithms use Piece-wise Quadratic potentials of SubQuadratic growth (PQSQ).
- PQSQ-based algorithms are as fast as the fast heuristic methods but more accurate.
- PQSQ-based algorithms are computationally efficient for regularized sparse regression.

ARTICLE INFO

Article history:
Received 26 May 2016
Received in revised form 10 August 2016
Accepted 19 August 2016
Available online 30 August 2016

Keywords:
Data approximation
Nonquadratic potential
Principal components
Clustering
Regularized regression
Sparse regression

ABSTRACT

Most of machine learning approaches have stemmed from the application of minimizing the mean squared distance principle, based on the computationally efficient quadratic optimization methods. However, when faced with high-dimensional and noisy data, the quadratic error functionals demonstrated many weaknesses including high sensitivity to contaminating factors and dimensionality curse. Therefore, a lot of recent applications in machine learning exploited properties of non-quadratic error functionals based on L_1 norm or even sub-linear potentials corresponding to quasinorms L_p ($0 < p < 1$). The back side of these approaches is increase in computational cost for optimization. Till so far, no approaches have been suggested to deal with *arbitrary* error functionals, in a flexible and computationally efficient framework. In this paper, we develop a theory and basic universal data approximation algorithms (k-means, principal components, principal manifolds and graphs, regularized and sparse regression), based on piece-wise quadratic error potentials of subquadratic growth (PQSQ potentials). We develop a new and universal framework to minimize *arbitrary sub-quadratic error potentials* using an algorithm with guaranteed fast convergence to the local or global error minimum. The theory of PQSQ potentials is based on the notion of the cone of minorant functions, and represents a natural approximation formalism based on the application of min-plus algebra. The approach can be applied in most of existing machine learning methods, including methods of data approximation and regularized and sparse regression, leading to the improvement in the computational cost/accuracy trade-off. We demonstrate that on synthetic and real-life datasets PQSQ-based machine learning methods achieve orders of magnitude faster computational performance than the corresponding state-of-the-art methods, having similar or better approximation accuracy.

1. Introduction

Modern machine learning and artificial intelligence methods are revolutionizing many fields of science today, such as medicine, biology, engineering, high-energy physics and sociology, where large amounts of data have been collected due to the emergence of new high-throughput computerized technologies. Historically and methodologically speaking, many machine learning algorithms have been based on minimizing the mean squared error potential, which can be explained by tractable properties of normal distribution and existence of computationally efficient methods for quadratic optimization. However, most of the real-life datasets

* Corresponding author.
E-mail addresses: ag153@le.ac.uk (A.N. Gorban), em322@le.ac.uk (E.M. Mirkes), Andrei.Zinovyev@curie.fr (A. Zinovyev).

http://dx.doi.org/10.1016/j.neunet.2016.08.007
0893-6080/

are characterized by strong noise, long-tailed distributions, presence of contaminating factors, large dimensions. Using quadratic potentials can be drastically compromised by all these circumstances: therefore, a lot of practical and theoretical efforts have been made in order to exploit the properties of non-quadratic error potentials which can be more appropriate in certain contexts. For example, methods of regularized and sparse regression such as lasso and elastic net based on the properties of L_1 metrics (Tibshirani, 1996; Zou & Hastie, 2005) found numerous applications in bioinformatics (Barillot, Calzone, Hupe, Vert, & Zinovyev, 2012), and L_1 norm-based methods of dimension reduction are of great use in automated image analysis (Wright et al., 2010). Not surprisingly, these approaches come with drastically increased computational cost, for example, connected with applying linear programming optimization techniques which are substantially more expensive compared to mean squared error-based methods.

In practical applications of machine learning, it would be very attractive to be able to deal with *arbitrary error potentials*, including those based on L_1 or fractional quasinorms L_p ($0 < p < 1$), in a computationally efficient and scalable way. There is a need in developing methods allowing to tune the *computational cost/accuracy of optimization* trade-off accordingly to various contexts.

In this paper, we suggest such a universal framework able to deal with a large family of error potentials. We exploit the fact that finding a minimum of a piece-wise quadratic function, or, in other words, a function which is the *minorant of a set of quadratic functionals*, can be almost as computationally efficient as optimizing the standard quadratic potential. Therefore, if a given arbitrary potential (such as L_1-based or fractional quasinorm-based) can be approximated by a piece-wise quadratic function, this should lead to relatively efficient and simple optimization algorithms. It appears that only potentials of quadratic or subquadratic growth are possible in this approach: however, these are the most useful ones in data analysis. We introduce a rich family of piece-wise quadratic potentials of subquadratic growth (PQSQ-potentials), suggest general approach for their optimization and prove convergence of a simple iterative algorithm in the most general case. We focus on the most used methods of data dimension reduction and regularized regression: however, potential applications of the approach can be much wider.

Data dimension reduction by constructing explicit low-dimensional approximators of a finite set of vectors is one of the most fundamental approach in data analysis. Starting from the classical data approximators such as k-means (Lloyd, 1957) and linear principal components (PCA) (Pearson, 1901), multiple generalizations have been suggested in the last decades (self-organizing maps, principal curves, principal manifolds, principal graphs, principal trees, etc.) (Gorban, Kegl, Wunsch, & Zinovyev, 2008; Gorban & Zinovyev, 2009) in order to make the data approximators more flexible and suitable for complex data structures.

We solve the problem of approximating a finite set of vectors $\vec{x}_i \in R^m$, $i = 1, \ldots, N$ (dataset) by a simpler object L embedded into the data space, such that for each point $\vec{x}_i$ an approximation error $err(\vec{x}_i, L)$ function can be defined. We assume this function in the form

$$err(\vec{x}_i, L) = \min_{y \in L} \sum_k u(x_i^k - y^k), \tag{1}$$

where the upper $k = 1, \ldots, m$ stands for the coordinate index, and $u(x)$ is a monotonously growing symmetrical function, which we will be calling the error potential. By data approximation we mean that the embedment of L in the data space minimizes the error

$$\sum_i err(\vec{x}_i, L) \to \min .$$

Note that our definition of error function is coordinate-wise (it is a sum of error potential over all coordinates).

The simplest form of the error potential is quadratic $u(x) = x^2$, which leads to the most known data approximators: mean point (L is a point), principal points (L is a set of points) (Flury, 1990), principal components (L is a line or a hyperplane) (Pearson, 1901). In more advanced cases, L can possess some regular properties leading to principal curves (L is a smooth line or spline) (Hastie, 1984), principal manifolds (L is a smooth low-dimensional surface) and principal graphs (eg., L is a pluri-harmonic graph embedment) (Gorban, Sumner, & Zinovyev, 2007; Gorban & Zinovyev, 2009).

There exist multiple advantages of using quadratic potential $u(x)$, because it leads to the most computationally efficient algorithms usually based on the splitting schema, a variant of expectation–minimization approach (Gorban & Zinovyev, 2009). For example, k-means algorithm solves the problem of finding the set of principal points and the standard iterative Singular Value Decomposition finds principal components. However, quadratic potential is known to be sensitive to outliers in the dataset. Also, purely quadratic potentials can suffer from the curse of dimensionality, not being able to robustly discriminate 'close' and 'distant' point neighbors in a high-dimensional space (Aggarwal, Hinneburg, & Keim, 2001).

There exist several widely used ideas for increasing approximator's robustness in presence of strong noise in data such as: (1) using medians instead of mean values, (2) substituting quadratic norm by L_1 norm (e.g. Ding, Zhou, He, & Zha, 2006 and Hauberg, Feragen, & Black, 2014), (3) outliers exclusion or fixed weighting or iterative reweighting during optimizing the data approximators (e.g. Allende, Rogel, Moreno, & Salas, 2004; Kohonen, 2001 and Xu & Yuille, 1995), and (4) regularizing the PCA vectors by L_1 norm (Candès, Li, Ma, & Wright, 2011; Jolliffe, Trendafilov, & Uddin, 2003; Zou, Hastie, & Tibshirani, 2006). In some works, it was suggested to utilize 'trimming' averages, e.g. in the context of the k-means clustering or some generalizations of PCA (Cuesta-Albertos, Gordaliza, & Matrán, 1997; Hauberg et al., 2014). In the context of regression, iterative reweighting is exploited to mimic the properties of L_1 norm (Lu, Lin, & Yan, 2015). Several algorithms for constructing PCA with L_1 norm have been suggested (Brooks, Dulá, & Boone, 2013; Ke & Kanade, 2005; Kwak, 2008) and systematically benchmarked (Brooks & Jot, 2012; Park & Klabjan, 2014). Some authors go even beyond linear metrics and suggest that fractional quasinorms L_p ($0 < p < 1$) can be more appropriate in high-dimensional data approximation (Aggarwal et al., 2001).

However, most of the suggested approaches exploiting properties of non-quadratic metrics either represent useful but still arbitrary heuristics or are not sufficiently scalable. The standard approach for minimizing L_1-based norm consists in solving a linear programming task. Despite existence of many efficient linear programming optimizer implementations, by their nature these computations are much slower than the iterative methods used in the standard SVD algorithm or k-means.

In this paper, we provide implementations of the standard data approximators (mean point, k-means, principal components) using a PQSQ potential. As an other application of PQSQ-based framework in machine learning, we develop PQSQ-based regularized and sparse regression (imitating the properties of lasso and elastic net).

2. Piecewise quadratic potential of subquadratic growth (PQSQ)

2.1. Definition of the PQSQ potential

Let us split all non-negative numbers $x \in R_{\geq 0}$ into $p + 1$ non-intersecting intervals $R_0 = [0; r_1), R_1 = [r_1; r_2), \ldots, R_k = [r_k; r_{k+1}), \ldots, R_p = [r_p; \infty)$, for a set of thresholds $r_1 < r_2 < \cdots < r_p$. For convenience, let us denote $r_0 = 0, r_{p+1} = \infty$. Piecewise quadratic potential is a continuous monotonously

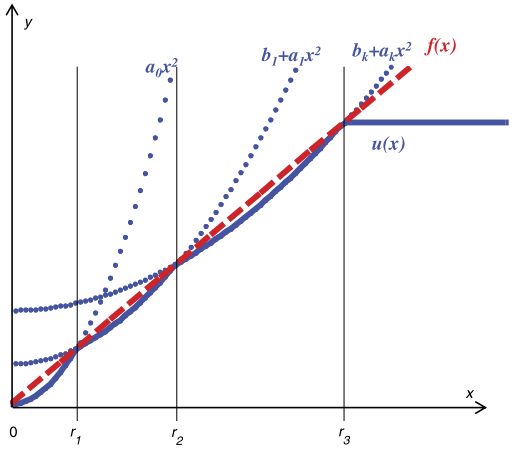

Fig. 1. Trimmed piecewise quadratic potential of subquadratic growth $u(x)$ (solid blue line) defined for the majorating function $f(x)$ (red dashed line) and several thresholds r_k. Dotted lines show the parabolas which fragments are used to construct $u(x)$. The last parabola is flat ($a_p = 0$) which corresponds to trimmed potential. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

growing function $u(x)$ constructed from pieces of centered at zero parabolas $y = b_k + a_k x^2$, defined on intervals $x \in [r_k, r_{k+1})$, satisfying $y(r_i) = f(r_i)$ (see Fig. 1):

$$u(x) = b_k + a_k x^2, \quad \text{if } r_k \le |x| < r_{k+1},\ k = 0, \ldots, p, \tag{2}$$

$$a_k = \frac{f(r_k) - f(r_{k+1})}{r_k^2 - r_{k+1}^2}, \tag{3}$$

$$b_k = \frac{f(r_{k+1})r_k^2 - f(r_k)r_{k+1}^2}{r_k^2 - r_{k+1}^2}, \tag{4}$$

where $f(x)$ is a majorating function, which is to be approximated (imitated) by $u(x)$. For example, in the simplest case $f(x)$ can be a linear function: $f(x) = x$, in this case, $\sum_k u(x^k)$ will approximate the L_1-based error function.

Note that accordingly to (3), (4), $b_0 = 0, a_p = 0, b_p = f(r_p)$. Therefore, the choice of r_p can naturally create a 'trimmed' version of error potential $u(x)$ such that some data points (outliers) do not have any contribution to the gradient of $u(x)$, hence, will not affect the optimization procedure. However, this set of points can change during minimization of the potential.

The condition of subquadratic growth consists in the requirement $a_{k+1} \le a_k$ and $b_{k+1} \ge b_k$. To guarantee this, the following simple condition on $f(x)$ should be satisfied:

$$f' > 0, \quad f''x \le f'. \tag{5}$$

Therefore, $f(x)$ is a monotonic concave function of $q = x^2$:

$$\frac{d^2 f(\sqrt{q})}{dq^2} = \frac{1}{4x^2} f''(x) - \frac{1}{4x^3} f'(x) \le 0.$$

In particular, $f(x)$ should grow not faster than any parabola $ax^2 + c$, $c > 0$, which is tangent to $f(x)$.

2.2. Basic approach for optimization

In order to use the PQSQ potential in an algorithm, a set of p interval thresholds r_s^k, $s = 1, \ldots, p$ for each coordinate $k = 1, \ldots, m$ should be provided. Matrices of a and b coefficients defined by (3) and (4) based on interval definitions: a_s^k, b_s^k, $s = 0, \ldots, p$, $k = 1, \ldots, m$ are computed separately for each coordinate k.

Minimization of PQSQ-based functional consists in several basic steps which can be combined in an algorithm:

(1) For each coordinate k, split all data point indices into non-overlapping sets $\mathcal{R}_s^k$:

$$\mathcal{R}_s^k = \{i : r_s^k \le |x_i^k - \beta_i^k| < r_{s+1}^k\}, \quad s = 0, \ldots, p, \tag{6}$$

where β is a matrix which depends on the nature of the algorithm.
(2) Minimize PQSQ-based functional where each set of points $\{x_{i \in \mathcal{R}_s^k}\}$ contributes to the functional quadratically with coefficient a_s^k. This is a quadratic optimization task.
(3) Repeat (1)–(2) till convergence.

3. General theory of the piece-wise convex potentials as the cone of minorant functions

In order to deal in most general terms with the data approximation algorithms based on PQSQ potentials, let us consider a general case where a potential can be constructed from a set of functions $\{q_i(x)\}$ with only two requirements: (1) that each $q_i(x)$ has a (local) minimum; (2) that the whole set of all possible $q_i(x)$s forms a cone. In this case, instead of the operational definition (2) it is convenient to define the potential $u(x)$ as the minorant function for a set of functions as follows. For convenience, in this section, x will notify a vector $\vec{x} \in R^m$.

Let us consider *a generating cone of functions* Q. We remind that the definition of a cone implies that for any $q(x) \in Q, p(x) \in Q$, we have $\alpha q(x) + \beta p(x) \in Q$, where $\alpha \ge 0, \beta \ge 0$.

For any finite set of functions

$$q_1(x) \in Q, q_2(x) \in Q, \ldots, q_s(x) \in Q,$$

we define the minorant function (Fig. 2):

$$u_{q_1,q_2,\ldots,q_s}(x) = \min(q_1(x), q_2(x), \ldots, q_s(x)). \tag{7}$$

It is convenient to introduce a multiindex

$$I_{q_1,q_2,\ldots,q_s}(x)$$

indicating which particular function(s) q_i corresponds to the value of $u(x)$, i.e.

$$I_{q_1,q_2,\ldots,q_s}(x) = \{i | u_{q_1,q_2,\ldots,q_s}(x) = q_i(x)\}. \tag{8}$$

For a cone Q let us define a set of all possible minorant functions $\mathbb{M}(Q)$

$$\mathbb{M}(Q) = \{u_{q_{i_1},q_{i_2},\ldots,q_{i_n}} | q_{i_1} \in Q, q_{i_2} \in Q, \quad q_{i_n} \in Q,\ n = 1, 2, 3, \ldots\}. \tag{9}$$

Proposition 1. *$\mathbb{M}(Q)$ is a cone.*

Proof. For any two minorant functions

$$u_{q_{i_1},q_{i_2},\ldots,q_{i_k}}, u_{q_{j_1},q_{j_2},\ldots,q_{j_s}} \in \mathbb{M}(Q)$$

we have

$$\alpha u_{q_{i_1},q_{i_2},\ldots,q_{i_k}} + \beta u_{q_{j_1},q_{j_2},\ldots,q_{j_s}} = u_{\{\alpha q_{i_p} + \beta q_{j_r}\}} \in \mathbb{M}(Q), \quad p = 1, \ldots, k,\ r = 1, \ldots, s, \tag{10}$$

where $\{\alpha q_{i_p} + \beta q_{j_r}\}$ is a set of all possible linear combinations of functions from $\{q_{i_1}, q_{i_2}, \ldots, q_{i_k}\}$ and $\{q_{j_1}, q_{j_2}, \ldots, q_{j_s}\}$.

Proposition 2. *Any restriction of $\mathbb{M}(Q)$ onto a linear manifold L is a cone.*

Proof. Let us denote $q(x)|_L$ a restriction of $q(x)$ function onto L, i.e. $q(x)|_L = \{q(x) | x \in L\}$. $q(x)|_L$ is a part of Q. Set of all $q(x)|_L$ forms a restriction $Q|_L$ of Q onto L. $Q|_L$ is a cone, hence, $\mathbb{M}(Q)|_L = \mathbb{M}(Q|_L)$ is a cone (Proposition 1).

Algorithm 1 Finding local minimum of a minorant function $u_{q_1,q_2,\ldots,q_n}(x)$

procedure MINIMIZING MINORANT FUNCTION
  initialize $x \leftarrow x_0$
repeat until stopping criterion has been met:
  compute multiindex $I_{q_1,q_2,\ldots,q_s}(x)$
  for all $i \in I_{q_1,q_2,\ldots,q_s}(x)$
  $x_i = \arg\min q_i(x)$
  end for
  select optimal x_i :
  $x_{opt} \leftarrow \arg\min_{x_i} u(x_i)$
  $x \leftarrow x_{opt}$
  stopping criterion: check if the multiindex $I_{q_1,q_2,\ldots,q_s}(x)$ does not change compared to the previous iteration
  end repeat:

Definition. *Splitting algorithm* minimizing

$u_{q_1,q_2,\ldots,q_n}(x)$

is defined as Algorithm 1.

Theorem 3.1. *Splitting algorithm (Algorithm 1) for minimizing* $u_{q_1,q_2,\ldots,q_n}(x)$ *converges in a finite number of steps.*

Proof. Since the set of functions $\{q_1, q_2, \ldots, q_n\}$ is finite then we only have to show that at each step the value of the function $u_{q_1,q_2,\ldots,q_n}(x)$ cannot increase. For any x and the value $x' = \arg\min q_i(x)$ for $i \in I_{q_1,q_2,\ldots,q_s}(x)$ we can have only two cases:

(1) Either $I_{q_1,q_2,\ldots,q_s}(x) = I_{q_1,q_2,\ldots,q_s}(x')$ (convergence, and in this case $q_{i'}(x') = q_i(x')$ for any $i' \in I_{q_1,q_2,\ldots,q_s}(x')$);

(2) Or $u_{q_1,q_2,\ldots,q_n}(x') < u_{q_1,q_2,\ldots,q_n}(x)$ since, accordingly to the definition (7), $q_{i'}(x') < q_i(x)$, for any $i' \in I_{q_1,q_2,\ldots,q_s}(x')$, $i \in I_{q_1,q_2,\ldots,q_s}(x)$ (see Fig. 2).

Note that in Algorithm 1 we do not specify exactly the way to find the local minimum of $q_i(x)$. To be practical, the cone Q should contain only functions for which finding a local minimum is fast and explicit. Evident candidates for this role are positively defined quadratic functionals $q(x) = q_0 + (\vec{q}_1, x) + (x, \mathbb{Q}_2 x)$, where $\mathbb{Q}_2$ is a positively defined symmetric matrix. Any minorant function (7) constructed from positively defined quadratic functions will automatically provide subquadratic growth, since the minorant cannot grow faster than any of the quadratic forms by which it is defined.

Operational definition of PQSQ given above (2), corresponds to a particular form of the quadratic functional, with $\mathbb{Q}_2$ being diagonal matrix. This choice corresponds to the coordinate-wise definition of data approximation error function (1) which is particularly simple to minimize. This circumstance is used in Algorithms 2, 3.

4. Commonly used data approximators with PQSQ potential

4.1. Mean value and k-means clustering in PQSQ approximation measure

Mean vector $\bar{X}_L$ for a set of vectors $X = \{x_i^k\}$, $i = 1, \ldots, N$, $k = 1, \ldots, m$ and an approximation error defined by potential $f(x)$ can be defined as a point minimizing the mean error potential for all points in X:

$$\sum_i \sum_k f(x_i^k - \bar{X}^k) \to \min . \tag{11}$$

For Euclidean metrics L_2 ($f(x) = x^2$) it is the usual arithmetic mean.

For L_1 metrics ($f(x) = |x|$), (11) leads to the implicit equation $\#(x_i^k > \bar{X}^k) = \#(x_i^k < \bar{X}^k)$, where # stands for the number of points, which corresponds to the definition of median. This equation can have a non-unique solution in case of even number of points or when some data point coordinates coincide: therefore, definition of median is usually accompanied by heuristics used for breaking ties, i.e. to deal with non-uniquely defined rankings. This situation reflects the general situation of existence of multiple local minimum and possible non-uniqueness of global minimum of (11) (Fig. 3).

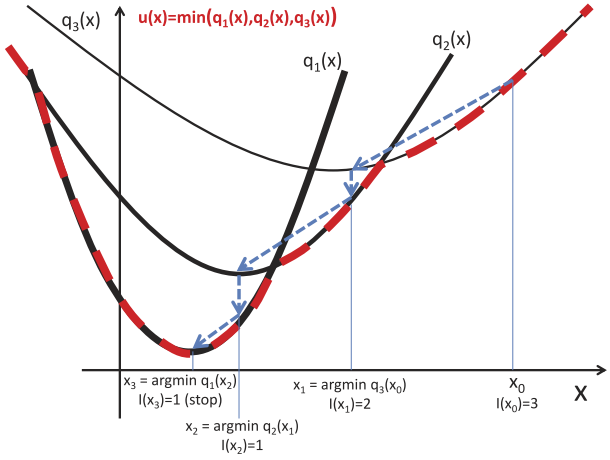

Fig. 2. Optimization of a one-dimensional minorant function $u(x)$, defined by three functions $q_1(x)$, $q_2(x)$, $q_3(x)$ each of which has a local minimum. Each optimization step consists in determining which $q_{I(x)}(x) = u(x)$ and making a step into the local minimum of $q_{I(x)}$.

For PQSQ approximation measure (2) it is difficult to write down an explicit formula for computing the mean value corresponding to the global minimum of (11). In order to find a point $\bar{X}_{PQSQ}$ minimizing mean *PQSQ* potential, a simple iterative algorithm can be used (Algorithm 2). The suggested algorithm converges to the local minimum which depends on the initial point approximation.

Algorithm 2 Computing PQSQ mean value

procedure PQSQ MEAN VALUE
  define intervals r_s^k, $s = 0, \ldots, p$, $k = 1, \ldots, m$
  compute coefficients a_s^k
  initialize $\bar{X}_{PQSQ}$
  eg., by arithmetic mean
repeat till convergence of $\bar{X}_{PQSQ}$:
  for each *coordinate* k
  define sets of indices

$$\mathcal{R}_s^k = \{i : r_s^k \le |x_i^k - \bar{X}_{PQSQ}^k| < r_{s+1}^k\}, \quad s = 0, \ldots, p$$

  compute new approximation for $\bar{X}_{PQSQ}$:

$$\bar{X}_{PQSQ}^k \leftarrow \frac{\sum_{s=1,\ldots,p} a_s^k \sum_{i \in \mathcal{R}_s^k} x_i^k}{\sum_{s=1,\ldots,p} a_s^k |\mathcal{R}_s^k|}$$

  end for
  goto *repeat till convergence*

Based on the PQSQ approximation measure and the algorithm for computing the PQSQ mean value (Algorithm 2), one can construct the PQSQ-based k-means clustering procedure in the usual way, splitting estimation of cluster centroids given partitioning of the data points into k disjoint groups, and then re-calculating the partitioning using the PQSQ-based proximity measure.

4.2. Principal Component Analysis (PCA) in PQSQ metrics

Accordingly to the classical definition of the first principal component, it is a line best fit to the dataset X (Pearson, 1901). Let

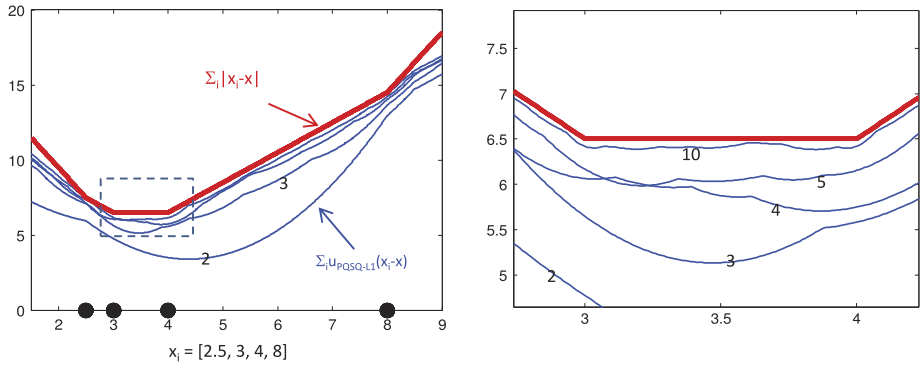

Fig. 3. Minimizing the error to a point (finding the mean value) for a set of 4 points (shown by black circles). Solid red line corresponds to L_1-based error. Thin blue lines correspond to PQSQ error potential imitating the L_1-based error. Several choices of PQSQ potential for different numbers of intervals (indicated by a number put on top of the line) is illustrated. On the right panel a zoom of a particular region of the left plot is shown. Neither function (L_1-based or PQSQ-based) possesses a unique local minimum. Moreover, L_1-based error function has infinite number of points corresponding to the global minimum (any number between 3 and 4), while PQSQ error function has several local minimum in [3;4] interval which exact positions are sensitive to the concrete choice of PQSQ parameters (interval definitions). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

us define a line in the parametric form $\vec{y} = \vec{V}v + \vec{\delta}$, where $v \in R^1$ is the parameter. Then the first principal component will be defined by vectors $\vec{V}, \vec{\delta}$ satisfying

$$\sum_i \sum_k u(x_i^k - V^k v_i - \delta^k) \to \min, \tag{12}$$

where

$$v_i = \arg\min_s \sum_k u(x_i^k - V^k s - \delta^k). \tag{13}$$

The standard first principal component (PC1) corresponds to $u(x) = x^2$ when the vectors $\vec{V}, \vec{\delta}$ can be found by a simple iterative splitting algorithm for Singular Value Decomposition (SVD). If X does not contain missing values then $\vec{\delta}$ is the vector of arithmetic mean values. By contrast, computing L_1-based principal components ($u(x) = |x|$) represents a much more challenging optimization problem (Brooks et al., 2013). Several approximative algorithms for computing L_1-norm PCA have been recently suggested and benchmarked (Brooks et al., 2013; Brooks & Jot, 2012; Ke & Kanade, 2005; Kwak, 2008; Park & Klabjan, 2014). To our knowledge, there have not been a general efficient algorithm suggested for computing PCA in case of arbitrary approximation measure for some monotonous function $u(x)$.

Computing PCA based on PQSQ approximation error is only slightly more complicated than computing the standard L_2 PCA by SVD. Here we provide a pseudo-code (Algorithm 3) of a simple iterative algorithm (similar to Algorithm 2) with guaranteed convergence (see Section 3).

Computation of second and further principal components follows the standard deflation approach: projections of data points onto the previously computed component are subtracted from the dataset, and the algorithm is applied to the residues. However, as it is the case in any non-quadratic metrics, the resulting components can be non-orthogonal or even not invariant with respect to the dataset rotation. Moreover, unlike L_2-based principal components, the Algorithm 3 does not always converge to a unique global minimum; the computed components can depend on the initial estimate of $\vec{V}$. The situation is somewhat similar to the standard k-means algorithm. Therefore, in order to achieve the least possible approximation error to the linear subspace, $\vec{V}$ can be initialized randomly or by data vectors $\vec{x}_i$ many times and the deepest in PQSQ approximation error (1) minimum should be selected.

How does the Algorithm 1 serve a more abstract version of the Algorithms 2, 3? For example, the 'variance' function $m(\vec{x}) = \frac{1}{N}\sum_j u(\vec{x}_j - \vec{x})$ to be minimized in Algorithm 2 uses the generating

Algorithm 3 Computing PQSQ PCA

procedure PQSQ First Principal Component
  define intervals r_s^k, $s = 0, \ldots, p$, $k = 1, \ldots, m$
  compute coefficients a_s^k
  $\vec{\delta} \leftarrow \vec{X}_{PQSQ}$
  initialize $\vec{V}$ *: eg., by* L_2*-based PC1*
  initialize $\{v_i\}$ *: eg., by*
  $v_i = \frac{\sum_k V^k(x_i^k - \delta^k)}{\sum_k (V^k)^2}$
repeat till convergence of $\vec{V}$*:*
  normalize $\vec{V}$: $\vec{V} \leftarrow \frac{\vec{V}}{\|\vec{V}\|}$
  for each *coordinate* k
  define sets of indices
  $\mathcal{R}_s^k = \{i : r_s^k \le |x_i^k - V^k v_i - \delta^k| < r_{s+1}^k\}$, $s = 0, \ldots, p$
  end for
  for each *data point* i and *coordinate* k
  find all $s_{i,k}$ *such that* $i \in \mathcal{R}_{s_{i,k}}^k$
  if *all* $a_{s_{i,k}}^k = 0$ ***then*** $v_i' \leftarrow 0$ **else**

$$v_i' \leftarrow \frac{\sum_k a_{s_{i,k}}^k V^k (x_i^k - \delta^k)}{\sum_k a_{s_{i,k}}^k (V^k)^2}$$

  end for
  for each *coordinate* k

$$V^k \leftarrow \frac{\sum_s a_s^k \sum_{i \in \mathcal{R}_s^k} (x_i^k - \delta^k) v_i}{\sum_s a_s^k \sum_{i \in \mathcal{R}_s^k} (v_i)^2}$$

  end for
  for each i :
  $v_i \leftarrow v_i'$
  end for
  goto *repeat till convergence*

functions in the form $Q = \{b_{ji}^k + \sum_k a_{ji}^k (x^k - x_j^k)^2\}$, where i is the index of the interval in (2). Hence, $m(x)$ is a minorant function, belonging to the cone $\mathbb{M}(Q)$, and must converge (to a local minimum) in a finite number of steps accordingly to Theorem 3.1.

4.3. Nonlinear methods: PQSQ-based principal graphs and manifolds

In a series of works, the authors of this article introduced a family of methods for constructing principal objects based on

graph approximations (e.g., principal curves, principal manifolds, principal trees), which allows constructing explicit non-linear data approximators (and, more generally, approximators with non-trivial topologies, suitable for approximating, e.g., datasets with branching or circular topology) (Gorban et al., 2008; Gorban' & Rossiev, 1999; Gorban et al., 2007; Gorban & Zinovyev, 2009, 2001a, 2001b, 2005, 2010). The methodology is based on optimizing a piece-wise quadratic *elastic energy* functional (see short description below). A convenient graphical user interface was developed with implementation of some of these methods (Gorban, Pitenko, & Zinovyev, 2014).

Let G be a simple undirected graph with set of vertices Y and set of edges E. For $k \geq 2$ a k-star in G is a subgraph with $k+1$ vertices $y_{0,1,\dots,k} \in Y$ and k edges $\{(y_0, y_i) \mid i = 1, \dots, k\} \subset E$. Suppose for each $k \geq 2$, a family S_k of k-stars in G has been selected. We call a graph G with selected families of k-stars S_k an *elastic graph* if, for all $E^{(i)} \in E$ and $S_k^{(j)} \in S_k$, the correspondent elasticity moduli $\lambda_i > 0$ and $\mu_{kj} > 0$ are defined. Let $E^{(i)}(0), E^{(i)}(1)$ be vertices of an edge $E^{(i)}$ and $S_k^{(j)}(0), \dots, S_k^{(j)}(k)$ be vertices of a k-star $S_k^{(j)}$ (among them, $S_k^{(j)}(0)$ is the central vertex).

For any map $\phi : Y \to R^m$ the *energy of the graph* is defined as

$$U^{\phi}(G) := \sum_{E^{(i)}} \lambda_i \left\| \phi(E^{(i)}(0)) - \phi(E^{(i)}(1)) \right\|^2 + \sum_{S_k^{(j)}} \mu_{kj} \left\| \sum_{i=1}^{k} \phi(S_k^{(j)}(i)) - k\phi(S_k^{(j)}(0)) \right\|^2.$$

For a given map $\phi : Y \to R^m$ we divide the dataset D into node neighborhoods K^y, $y \in Y$. The set K^y contains the data points for which the node $\phi(y)$ is the closest one in ϕ. The *energy of approximation* is:

$$U_A^{\phi}(G, D) = \sum_{y \in Y} \sum_{x \in K^y} w(x) \|x - \phi(y)\|^2, \tag{14}$$

where $w(x) \geq 0$ are the point weights. Simple and fast algorithm for minimization of the energy

$$U^{\phi} = U_A^{\phi}(G, D) + U^{\phi}(G) \tag{15}$$

is the splitting algorithm, in the spirit of the classical k-means clustering: for a given system of sets $\{K^y \mid y \in Y\}$ we minimize U^{ϕ} (optimization step, it is the minimization of a positive quadratic functional), then for a given ϕ we find new $\{K^y\}$ (re-partitioning), and so on; stop when no change.

Application of PQSQ-based potential is straightforward in this approach. It consists in replacing (14) with

$$U_A^{\phi}(G, D) = \sum_{y \in Y} \sum_{x \in K^y} w(x) \sum_k u(x^k - \phi(y^k)),$$

where u is a chosen PQSQ-based error potential. Partitioning of the dataset into $\{K^y\}$ can be also based on calculating the minimum PQSQ-based error to y, or can continue enjoying nice properties of L_2-based distance calculation.

5. PQSQ-based regularized regression

5.1. Regularizing linear regression with PQSQ potential

One of the major application of non-Euclidean norm properties in machine learning is using non-quadratic terms for penalizing large absolute values of regression coefficients (Tibshirani, 1996; Zou & Hastie, 2005). Depending on the chosen penalization term, it is possible to achieve various effects such as sparsity or grouping coefficients for redundant variables. In a general form, regularized regression solves the following optimization problem

$$\frac{1}{N} \sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{m} \beta^k x_i^k \right)^2 + \lambda f(\vec{\beta}) \to \min, \tag{16}$$

where N is the number of observations, m is the number of independent variables in the matrix $\{x_i^k\}$, $\{y_i\}$ are values of the dependent variable (to be predicted), λ is an internal parameter controlling the strength of regularization (penalty on the amplitude of regression coefficients β), and $f(\vec{z})$ is the regularizer function, which is $f(\vec{z}) = \|\vec{z}\|_{L2}^2$ for ridge regression, $f(\vec{z}) = \|\vec{z}\|_{L1}$ for lasso and $f(\vec{z}) = \frac{1-\alpha}{\alpha}\|\vec{z}\|_{L2}^2 + \alpha\|\vec{z}\|_{L1}$ for elastic net methods correspondingly.

Here we suggest to imitate $f(x)$ with a *PQSQ* potential function, i.e. instead of (16) solving the problem

$$\frac{1}{N} \sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{m} \beta^k x_i^k \right)^2 + \lambda \sum_{k=1}^{m} u(\beta^k) \to \min, \tag{17}$$

where $u(\beta)$ is a PQSQ potential imitating *arbitrary* subquadratic regression regularization penalty.

Solving (17) is equivalent to iteratively solving a system of linear equations

$$\frac{1}{N} \sum_{k=1}^{m} \beta^k \sum_{i=1}^{N} x_i^k x_i^j + \lambda a_{I(\beta^j)} \beta^j = \sum_{i=1}^{N} y_i x_i^j, \quad j = 1, \dots, m, \tag{18}$$

where $a_{I(\beta^j)}$ constant (where I index is defined from $r_I \leq \beta^j < r_{I+1}$) is computed accordingly to the definition of $u(x)$ function (see (3)), given the estimation of β^k regression coefficients at the current iteration. In practice, iterating (18) converges in a few iterations, therefore, the algorithm can work very fast and outperform the widely used least angle regression algorithm for solving (16) in case of L_1 penalties.

5.2. Introducing sparsity by 'black hole' trick

Any PQSQ potential $u(x)$ is characterized by zero derivative for $x = 0$ by construction: $u'(x)|_{x=0} = 0$, which means that the solution of (17) does not have to be sparse for any λ. Unlike pure L_1-based penalty, the coefficients of regression diminish with increase of λ but there is nothing to shrink them to exact zero values, similar to the ridge regression. However, it is relatively straightforward to modify the algorithm, to achieve sparsity of the regression solution. The 'black hole' trick consists in eliminating from regression training after each iteration (18) all regression coefficients β^k smaller by absolute value than a given parameter ϵ ('black hole radius'). Those regression coefficients which have passed the 'black hole radius' are put to zero and do not have any chance to change their values in the next iterations.

The optimal choice of ϵ value requires a separate study. From general considerations, it is preferable that the derivative $u'(x)|_{x=\epsilon}$ would not be very close to zero. As a pragmatic choice for the numerical examples in this article, we define ϵ as the midst of the smallest interval in the definition of PQSQ potential (see Fig. 4), i.e. $\epsilon = r_1/2$, which guarantees far from zero $u'(x)|_{x=\epsilon}$. It might happen that this value of ϵ would collapse all β^k to zero even without regularization (i.e., with $\lambda = 0$). In this case, the 'black hole radius' is divided by half $\epsilon \leftarrow \epsilon/2$ and it is checked that for $\lambda = 0$ the iterations would leave at list half of the regression coefficients. If it is not the case then the process of diminishing the 'black hole radius' repeated recursively till meeting the criterion of preserving

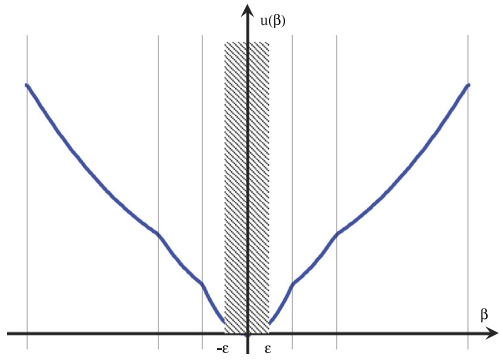

Fig. 4. 'Black hole trick' for introducing sparsity into the PQSQ-based regularized regression. Here PQSQ function imitates L_1 norm (for illustration only three intervals are used to define PQSQ function). Black hole trick consists in introducing an ϵ zone (hatched territory on the plot) of the potential in the vicinity of zero such that any coefficient of regression falling down into this zone is set to zero and eliminated from further learning. It is convenient to define ϵ as the midst of the smallest interval as it is shown in this plot.

the majority of regression coefficients. In practice, it requires only few (very fast) additional iterations of the algorithm.

As in the lasso methodology, the problem (17) is solved for a range of λ values, calibrated such that the minimal λ would select the maximum number of regression variables, while the maximum λ value would lead to the most sparse regression (selecting only one single non-zero regression coefficient).

6. Numerical examples

6.1. Practical choices of parameters

The main parameters of PQSQ are (a) majorating function $f(x)$ and (b) decomposition of each coordinate range into $p+1$ non-overlapping intervals. Depending on these parameters, various approximation error properties can be exploited, including those providing robustness to outlier data points.

When defining the intervals r_j, $j = 1, \dots, p$, it is desirable to achieve a small difference between $f(\Delta x)$ and $u(\Delta x)$ for expected argument values Δx (differences between an estimator and the data point), and choose the suitable value of the potential trimming threshold r_p in order to achieve the desired robustness properties. If no trimming is needed, then r_p should be made larger than the maximum expected difference between coordinate values (maximum Δx).

In our numerical experiments we used the following definition of intervals. For any data coordinate k, we define a characteristic difference D^k, for example

$$D^k = \alpha_{scale}(\max_i(x_i^k) - \min_i(x_i^k)), \tag{19}$$

where α_{scale} is a scaling parameter, which can be put at 1 (in this case, the approximating potential will not be trimmed). In case of existence of outliers, for defining D^k, instead of amplitude one can use other measures such as the median absolute deviation (MAD):

$$D^k = \alpha_{scale} median_i(|x_i^k - median(\{x_i^k\})|); \tag{20}$$

in this case, the scaling parameter should be made larger, i.e. $\alpha_{scale} = 10$, if no trimming is needed.

After defining D^k we use the following definition of intervals:

$$r_j^k = D^k \frac{j^2}{p^2}, \quad j = 0, \dots, p. \tag{21}$$

More sophisticated approaches are also possible to apply such as, given the number of intervals p and the majorant function $f(x)$, choose r_j, $j = 1, \dots, p$ in order to minimize the maximal difference

$$d = \max_x |f(x) - u(x)| \to \min.$$

The calculation of intervals is straightforward for a given value of d and many smooth concave functions $f(x)$ like $f(x) = x^p$ $(0 < p \leq 1)$ or $f(x) = \ln(1+x)$.

6.2. Implementation

We provide Matlab implementation of PQSQ approximators (in particular, PCA) together with the Matlab and R code used to generate the example figures in this article at 'PQSQ-DataApproximators' GitHub repository[1] and Matlab implementation of PQSQ-based regularized regression with build-in imitations of L_1 (lasso-like) and $L_1\&L_2$ mixture (elastic net-like) penalties at 'PQSQ-regularized-regression' GitHub repository.[2] The Java code implementing elastic graph-based non-linear approximator implementations is available from the authors on request.

6.3. Motivating example: dense two-cluster distribution contaminated by sparse noise

We demonstrate the value of PQSQ-based computation of L_1-based PCA by constructing a simple example of data distribution consisting of a dense two-cluster component superimposed with a sparse contaminating component with relatively large variance whose co-variance does not coincide with the dense signal (Fig. 5). We study the ability of PCA to withstand certain level of sparse contamination and compare it with the standard L_2-based PCA. In this example, without noise the first principal component coincides with the vector connecting the two cluster centers: hence, it perfectly separates them in the projected distribution. Noise interferes with the ability of the first principal component to separate the clusters to the degree when the first principal component starts to match the principal variance direction of the contaminating distribution (Fig. 5(A), (B)). In higher dimensions, not only the first but also the first two principal components are not able to distinguish two clusters, which can hide an important data structure when applying the standard data visualization tools.

In the first test we study a switch of the first principal component from following the variance of the dense informative distribution (abscissa) to the sparse noise distribution (ordinate) as a function of the number of contaminating points, in R^2 (Fig. 5(A)–(C)). We modeled two clusters as two 100-point samples from normal distribution centered in points $[-1; 0]$ and $[1; 0]$ with isotropic variance with the standard deviation 0.1. The sparse noise distribution was modeled as a k-point sample from the product of two Laplace distributions of zero means with the standard deviations 2 along abscissa and 4 along ordinate. The intervals for computing the PQSQ functional were defined by thresholds $R = \{0; 0.01; 0.1; 0.5; 1\}$ for each coordinate. Increasing the number of points in the contaminating distribution diminishes the average value of the abscissa coordinate of PC1, because the PC1 starts to be attracted by the contaminating distribution (Fig. 5(C)). However, it is clear that on average PQSQ L_1-based PCA is able to withstand much larger amplitude of the contaminating signal (very robust up to 20–30 points, i.e. 10%–20% of strong noise contamination) compared to the standard L_2-based PCA (which is robust to 2%–3% of contamination).

In the second test we study the ability of the first two principal components to separate two clusters, in R^{100} (Fig. 5(D)–(F)). As in

[1] https://github.com/auranic/PQSQ-DataApproximators.
[2] https://github.com/Mirkes/PQSQ-regularized-regression/.

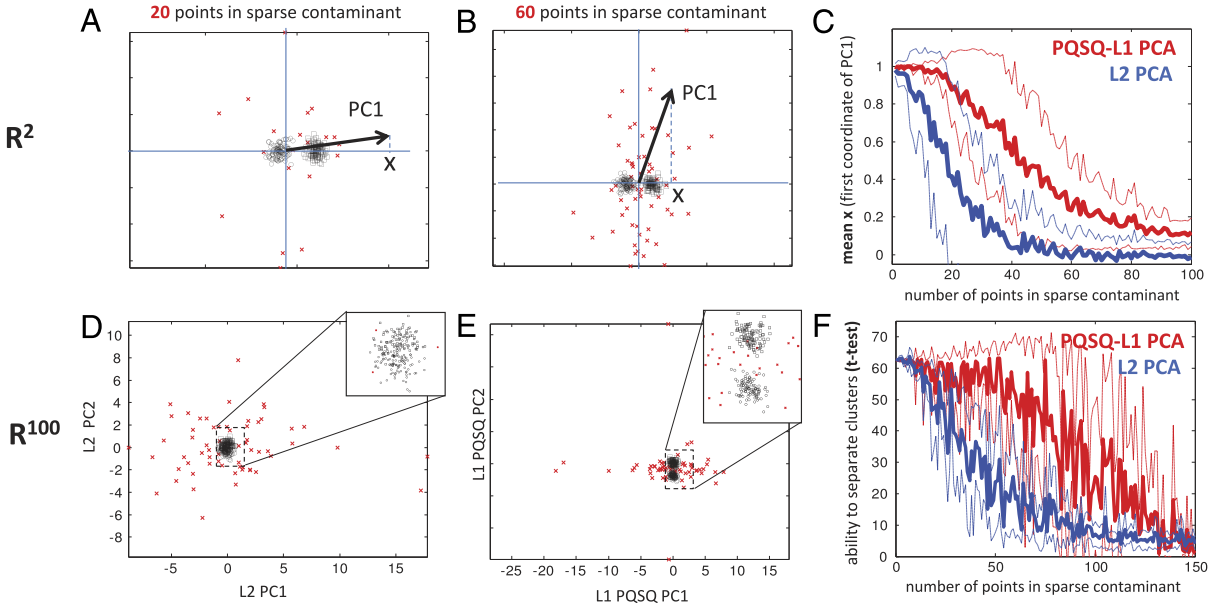

Fig. 5. Comparing L_2- and PQSQ L_1-based PCA using example of two-cluster distribution (100 black circles and 100 squares) contaminated by sparse noise (red crosses). (A) Dense two cluster distribution contaminated by sparse distribution (20 points) of large variance. In the presence of noise, the abscissa coordinate x of PC1 Vector is slightly less than 1. (B) Same as (A) but in the case of strong contamination (60 points). The value of x is much smaller in this case. (C) Average absolute value of the abscissa coordinate of PC1 $|x|$ (thick lines) shown with standard interval (thin lines) for 100 samples of k contaminating points. (D) Projection of the data distribution on the first two principal components computed with the standard L_2 PCA algorithm. The number of contaminating points is 40. The cluster structure of the dense part of the distribution is completely hidden as shown in the zoom window. (E) Same as in (D) but computed with PQSQ L_1-based algorithm. The cluster structure is perfectly separable. (F) The value of t-test computed based on the known cluster labels of the dense part of the distribution, in the projections onto the first two principal components of the global distribution. As in (C), the mean values of 100 contamination samples together with confidence intervals are shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the first test, we modeled two clusters as two 100-point samples from normal distribution centered in points $[-1; 0; \ldots; 0]$ and $[1; 0; \ldots; 0]$ with isotropic variance with the standard deviation 0.1 in all 100 dimensions. The sparse noise distribution is modeled as a k-point sample from the product of 100 Laplace distributions of zero means with the standard deviations 1 along each coordinate besides the third coordinate (standard deviation of noise is 2) and the fourth coordinate (standard deviation of noise is 4). Therefore, the first two principal components in the absence of noise are attracted by the dimension 1 and noise, while in the presence of strong noise they are be attracted by dimensions 3 and 4, hiding the cluster structure of the dense part of the distribution. The definitions of the intervals were taken as in the first test. We measured the ability of the first two principal components to separate clusters by computing the t-test between the two clusters projected in the 2D-space spanned by the first principal components of the global distribution (Fig. 5(D)–(F)). As one can see, the average ability of the first principal components to separate clusters is significantly stronger in the case of PQSQ L_1-based PCA which is able to separate robustly the clusters even in the presence of strong noise contamination (up to 80 noise points, i.e. 40% contamination).

6.4. Performance/stability trade-off benchmarking of L_1-based PCA

In order to compare the computation time and the robustness of PQSQ-based PCA algorithm for the case $u(x) = |x|$ with existing R-based implementations of L_1-based PCA methods (pcaL1 package), we followed the benchmark described in Brooks and Jot (2012). We compared performance of PQSQ-based PCA based on Algorithm 3 with several L_1-based PCA algorithms: L1-PCA* (Brooks et al., 2013), L1-PCA (Ke & Kanade, 2005), PCA-PP (Croux, Filzmoser, & Oliveira, 2007), PCA-L1 (Kwak, 2008). As a reference point, we used the standard PCA algorithm based on quadratic norm and computed using the standard SVD iterations.

The idea of benchmarking is to generate a series of datasets of the same size ($N = 1000$ objects in $m = 10$ dimensions) such that the first 5 dimensions would be sampled from a uniform distribution $U(-10, 10)$. Therefore, the first 5 dimensions represent 'true manifold' sampled by points.

The values in the last 5 dimensions represent 'noise+outlier' signal. The background noise is represented by Laplacian distribution of zero mean and 0.1 variance. The outlier signal is characterized by mean value μ, dimension p and frequency ϕ. Then, for each data point with a probability ϕ, in the first p outlier dimensions a value is drawn from $Laplace(\mu, 0.1)$. The rest of the values is drawn from background noise distribution.

As in Brooks and Jot (2012), we have generated 1300 test sets corresponding to $\phi = 0.1$, with 100 examples for each combination of $\mu = 1, 5, 10, 25$ and $p = 1, 2, 3$.

For each test set 5 first principal components $\vec{V}_1, \ldots, \vec{V}_5$ of unit length were computed, with corresponding point projection distributions $U^1, \ldots, U^5$ and the mean vector $\vec{C}$. Therefore, for each initial data point $\vec{x}_i$, we have the 'restored' data point

$$P(\vec{x}_i) = \sum_{k=1,\ldots,5} U_i^k \vec{V}_k + \vec{C}.$$

For computing the PQSQ-based PCA we used 5 intervals without trimming. Changing the number of intervals did not significantly changed the benchmarking results.

Two characteristics were measured: (1) computation time measured as a ratio to the computation of 5 first principal components using the standard L_2-based PCA and (2) the sum of absolute values of the restored point coordinates in the 'outlier' dimensions normalized on the number of points:

$$\sigma = \frac{1}{N} \sum_{i=1,\ldots,N} \sum_{k=6,\ldots,10} |P^k(\vec{x}_i)|. \tag{22}$$

Formally speaking, σ is L_1-based distance from the point projection onto the first five principal components to the 'true'

Table 1
Comparing time performance (in seconds, on ordinary laptop) of lasso vs. PQSQ-based regularized regression imitating L_1 penalty. Average acceleration of PQSQ-based method vs. lasso in these 8 examples is 120 fold with comparable accuracy.

Dataset	Objects	Variables	lasso	PQSQ	Ratio
Breast cancer	47	31	10.50	0.05	233.33
Prostate cancer	97	8	0.07	0.02	4.19
ENB2012	768	8	0.53	0.03	19.63
Parkinson	5875	26	20.30	0.04	548.65
Crime	1994	100	10.50	0.19	56.24
Crime reduced	200	100	17.50	0.17	102.94
Forest fires	517	8	0.05	0.02	3.06
Random regression (1000 × 250)	1000	250	2.82	0.58	4.86

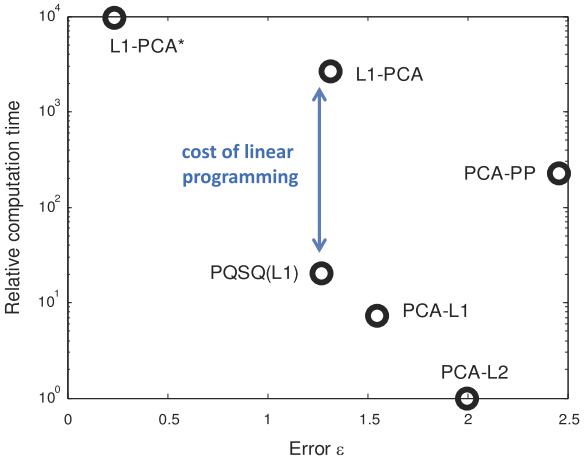

Fig. 6. Benchmarking several algorithms for constructing L_1-based PCA, using synthetic datasets representing 'true' five-dimensional linear manifold contaminated by noise and outliers (located in other five dimensions). The abscissa is the error of detecting the 'true' manifold by a particular method and the ordinate is the computational time relative to the standard SVD (L_2-based PCA) computation, in logarithmic scale. The computational cost of application of linear programming methods instead of simpler iterative methods is approximately shown by an arrow.

subspace. In simplistic terms, larger values of σ correspond to the situation when the first five principal components do not represent well the first 'true' dimensions having significant loadings into the 'outlier dimensions'. $\sigma = 0$ if and only if the first five components do not have any non-zero loadings in the dimensions 6, . . . , 10.

The results of the comparison averaged over all 1300 test sets are shown in Fig. 6. The PQSQ-based computation of PCA outperforms by accuracy the existing heuristics such as PCA-L1 but remains computationally efficient. It is 100 times faster than L1-PCA giving almost the same accuracy. It is almost 500 times faster than the L1-PCA* algorithm, which is, however, better in accuracy (due to being robust with respect to strong outliers). One can see from Fig. 6 that PQSQ-based approach is the best in accuracy among fast iterative methods. The detailed tables of comparison for all combinations of parameters are available on GitHub.[3] The scripts used to generate the datasets and compare the results can also be found there.[4]

6.5. Comparing performances of PQSQ-based regularized regression and lasso algorithms

We compared performance of PQSQ-based regularized regression method imitating L_1 penalty with lasso implementation in Matlab, using 8 datasets from UCI Machine Learning Repository (Lichman, 2013), Regression Task section. In the selection of datasets we chose very different numbers of objects and variables for regression construction (Table 1). All table rows containing missing values were eliminated for the tests.

We observed up to two orders of magnitude acceleration of PQSQ-based method compared to the lasso method implemented in Matlab (Table 1), with similar sparsity properties and approximation power as lasso (Fig. 7).

While comparing time performances of two methods, we have noticed that lasso (as it is implemented in Matlab) showed worse results when the number of objects in the dataset approaches the number of predictive variables (see Table 1). In order to test this observation explicitly, we took a particular dataset ('Crime') containing 1994 observations and 100 variables and compared the performance of lasso in the case of complete table and a reduced table ('Crime reduced') containing only each 10th observation. Paradoxically, lasso converges almost two times slower in the case of the smaller dataset, while the PQSQ-based algorithm worked slightly faster in this case.

It is necessary to stress that here we compare the basic algorithms without many latest technical improvements which can be applied both to L_1 penalty and its PQSQ approximation (such as fitting the whole lasso path). Detailed comparison of all the existent modifications if far beyond the scope of this work.

For comparing approximation power of the PQSQ-based regularized regression and lasso, we used two versions of PQSQ potential for regression coefficients: with and without trimming. In order to represent the results, we used the 'Number of non-zero parameters vs. Fraction of Variance Unexplained (FVU)' plots (see two representative examples at Fig. 7). We suggest that this type of plot is more informative in practical applications than the 'lasso plot' used to calibrate the strength of regularization, since it is a more explicit representation for optimizing the accuracy vs. complexity ratio of the resulting regression.

From our testing, we can conclude that PQSQ-based regularized regression has similar properties of sparsity and approximation accuracy compared to lasso. It tends to slightly outperform lasso (to give smaller FVU) in case of $N \approx P$. Introducing trimming in most cases does not change the best FVU for a given number of selected variables, but tends to decrease its variance (has a stabilization effect). In some cases, introducing trimming is the most advantageous method (Fig. 7(B)).

The GitHub 'PQSQ-regularized-regression' repository contains exact dataset references and more complete report on comparing approximation ability of PQSQ-based regularized regression with lasso.[5]

7. Conclusion

In this paper we develop a new machine learning framework (theory and application) allowing one to deal with arbitrary error potentials of not-faster than quadratic growth, imitated by

[3] http://goo.gl/sXBvqh.
[4] https://github.com/auranic/PQSQ-DataApproximators.
[5] https://github.com/Mirkes/PQSQ-regularized-regression/wiki.

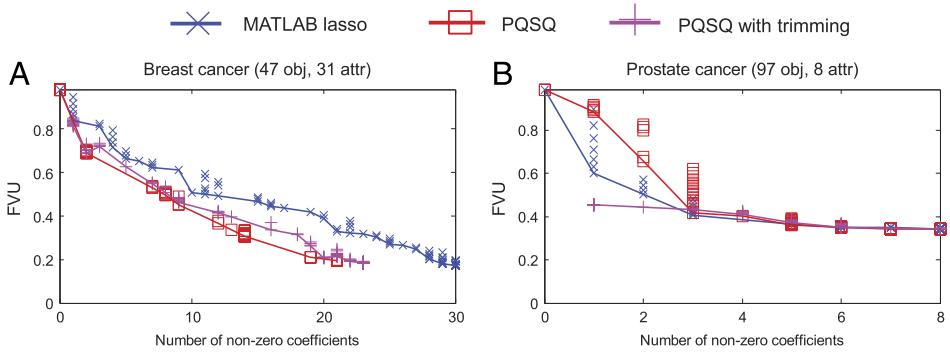

Fig. 7. Number of non-zero regression coefficients vs. FVU plot for two example real-life datasets (A–Breast cancer Wisconsin dataset from UC Irvine Machine Learning Repository, B—original prostate cancer example from the seminal lasso paper Tibshirani, 1996). Each cross shows a particular solution of the regularized regression problem. Solid lines show the best (minimal) obtained FVU within the same number of selected variables.

Table 2
List of methods which can use PQSQ-based error potentials.

Data approximation/Clustering/Manifold learning	
Principal component analysis	Includes robust trimmed version of PCA, L_1-based PCA, regularized PCA, and many other PCA modifications
Principal curves and manifolds	Provides possibility to use non-quadratic data approximation terms and trimmed robust version
Self-organizing maps	Same as above
Principal graphs/trees	Same as above
k-means	Can include adaptive error potentials based on estimated error distributions inside clusters
High-dimensional data mining	
Use of fractional quasinorms L_p $(0 < p < 1)$	Introducing fractional quasinorms in existing data-mining techniques can potentially deal with the curse of dimensionality, helping to better distinguish close from distant data points (Aggarwal et al., 2001)
Regularized and sparse regression	
Lasso	Application of PQSQ-based potentials leads to speeding up in case of large and $N \approx P$ datasets
Elastic net	Same as above

piece-wise quadratic function of subquadratic growth (PQSQ error potential).

We develop methods for constructing the standard data approximators (mean value, k-means clustering, principal components, principal graphs) for arbitrary non-quadratic approximation error with subquadratic growth and regularized linear regression with arbitrary subquadratic penalty by using a piecewise-quadratic error functional (PQSQ potential). These problems can be solved by applying quasi-quadratic optimization procedures, which are organized as solutions of sequences of linear problems by standard and computationally efficient algorithms.

The suggested methodology have several advantages over existing ones:

(a) *Scalability*: the algorithms are computationally efficient and can be applied to large data sets containing millions of numerical values.
(b) *Flexibility*: the algorithms can be adapted to any type of data metrics with subquadratic growth, even if the metrics cannot be expressed in explicit form. For example, the error potential can be chosen as adaptive metrics (Wu, Jin, Hoi, Zhu, & Yu, 2009; Yang & Jin, 2006).
(c) *Built-in (trimmed) robustness*: choice of intervals in PQSQ can be done in the way to achieve a trimmed version of the standard data approximators, when points distant from the approximator do not affect the error minimization during the current optimization step.
(d) *Guaranteed convergence*: the suggested algorithms converge to local or global minimum just as the corresponding predecessor algorithms based on quadratic optimization and expectation/minimization-based splitting approach.

In theoretical perspective, using PQSQ-potentials in data mining is similar to existing applications of min-plus (or, max-plus) algebras in non-linear optimization theory, where complex non-linear functions are approximated by infimum (or supremum) of finitely many 'dictionary functions' (Gaubert, McEneaney, & Qu, 2011; Magron, Allamigeon, Gaubert, & Werner, 2015). We can claim that just as using polynomes is a natural framework for approximating in rings of functions, using min-plus algebra naturally leads to introduction of PQSQ-based functions and the cones of minorants of quadratic dictionary functions.

One of the application of the suggested methodology is approximating the popular in data mining L_1 metrics. We show by numerical simulations that PQSQ-based approximators perform as fast as the fast heuristical algorithms for computing L_1-based PCA but achieve better accuracy in a previously suggested benchmark test. PQSQ-based approximators can be less accurate than the exact algorithms for optimizing L_1-based functions utilizing linear programming; however, they are several orders of magnitude faster. PQSQ potential can be applied in the task of regression, replacing the classical Least Squares or L_1-based Least Absolute Deviation methods. At the same time, PQSQ-based approximators can imitate a variety of subquadratic error potentials (not limited to L_1 or variations), including fractional quasinorms L_p $(0 < p < 1)$. We demonstrate that the PQSQ potential can be easily adapted to the problems of sparse regularized regression with non-quadratic penalty on regression coefficients (including imitations of lasso and elastic net). On several real-life dataset examples we show that PQSQ-based regularized regression can perform two orders of magnitude faster than the lasso algorithm implemented in the same programming environment.

To conclude, in Table 2 we list possible applications of the PQSQ-based framework in machine learning.

Acknowledgment

This study was supported in part by Big Data Paris Science et Lettre Research University project 'PSL Institute for Data Science'.

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *Database theory - ICDT 2001, 8th international conference* (pp. 420–434). Springer.

Allende, H., Rogel, C., Moreno, S., & Salas, R. (2004). Robust neural gas for the analysis of data with outliers. In *Computer science society. SCCC 2004. 24th International conference of the Chilean* (pp. 149–155). IEEE.

Barillot, E., Calzone, L., Hupe, P., Vert, J.-P., & Zinovyev, A. (2012). *CRC mathemtical and computational biology, Computational systems biology of cancer*. Chapman & Hall.

Brooks, J., Dulá, J., & Boone, E. (2013). A pure L1-norm principal component analysis. *Computational Statistics & Data Analysis*, *61*, 83–98.

Brooks, J., & Jot, S. (2012). PCAL1: An implementation in R of three methods for L1-norm principal component analysis, Optimization Online preprint, http://www.optimization-online.org/DB_HTML/2012/04/3436.html.

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *Journal of the ACM*, *58*, 11.

Croux, C., Filzmoser, P., & Oliveira, M. R. (2007). Algorithms for projection–pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *87*, 218–225.

Cuesta-Albertos, J., Gordaliza, A., Matrán, C., et al. (1997). Trimmed *k*-means: An attempt to robustify quantizers. *The Annals of Statistics*, *25*, 553–576.

Ding, C., Zhou, D., He, X., & Zha, H. (2006). R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization. In *ICML* (pp. 281–288).

Flury, B. (1990). Principal points. *Biometrika*, *77*, 33–41.

Gaubert, S., McEneaney, W., & Qu, Z. (2011). Curse of dimensionality reduction in max-plus based approximation methods: theoretical estimates and improved pruning algorithms, Arxiv Preprint http://arxiv.org/abs/1109.5241.

Gorban, A., Kegl, B., Wunsch, D., & Zinovyev, A. (Eds.) (2008). *LNCSE: Vol. 58. Principal manifolds for data visualisation and dimension reduction*. Springer.

Gorban, A.N., Pitenko, A., & Zinovyev, A. (2014). ViDaExpert: user-friendly tool for nonlinear visualization and analysis of multidimensional vectorial data, ArXiv Preprint http://arxiv.org/abs/1406.5550.

Gorban', A., & Rossiev, A. (1999). Neural network iterative method of principal curves for data with gaps. *Journal of Computer and Systems Sciences International c/c of Tekhnicheskaia Kibernetika*, *38*, 825–830.

Gorban, A. N., Sumner, N. R., & Zinovyev, A. Y. (2007). Topological grammars for data approximation. *Applied Mathematics Letters*, *20*, 382–386.

Gorban, A. N., & Zinovyev, A. (2009). Principal graphs and manifolds. In E. S. Olivas, J. D. M. Guererro, M. M. Sober, J. R. M. Benedito, & A. J. S. Lopes (Eds.), *Handbook of research on machine learning applications and trends: Algorithms, methods and techniques*.

Gorban, A., & Zinovyev, A. (2001a). Visualization of data by method of elastic maps and its applications in genomics, economics and sociology, IHES Preprints.

Gorban, A., & Zinovyev, A. Y. (2001b). Method of elastic maps and its applications in data visualization and data modeling. *International Journal of Computing Anticipatory Systems, CHAOS*, *12*, 353–369.

Gorban, A., & Zinovyev, A. (2005). Elastic principal graphs and manifolds and their practical applications. *Computing*, *75*, 359–379.

Gorban, A. N., & Zinovyev, A. (2010). Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *International Journal of Neural Systems*, *20*, 219–232.

Hastie, T. (1984). *Principal curves and surfaces*. (Ph.D. Thesis), California: Stanford University.

Hauberg, S., Feragen, A., & Black, M.J. (2014). Grassmann averages for scalable robust pca. In *2014 IEEE conference on computer vision and pattern recognition* (pp. 3810–3817).

Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, *12*, 531–547.

Ke, Q., & Kanade, T. (2005). Robust l 1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *IEEE computer society conference on Computer vision and pattern recognition: Vol. 1* (pp. 739–746). IEEE.

Kohonen, T. (2001). *Springer series in information sciences: Vol. 30. Self-organizing maps*. Berlin: Springer.

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

Lichman, M. (2013). University of California, Irvine (UCI) Machine Learning Repository, http://archive.ics.uci.edu/ml.

Lloyd, S. (1957). Last square quantization in pcm's. *Bell Telephone Laboratories Paper*,.

Lu, C., Lin, Z., & Yan, S. (2015). Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization. *IEEE Transactions on Image Processing*, *24*, 646–654.

Magron, V., Allamigeon, X., Gaubert, S., & Werner, B. (2015). Formal proofs for nonlinear optimization, Arxiv Preprint arXiv:1404.7282.

Park, Y.W., & Klabjan, D. (2014). Algorithms for L1-norm principal component analysis. Tutorial, http://dynresmanagement.com/uploads/3/3/2/9/3329212/algorithms_for_l1pca.pdf.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*, 559–572.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 267–288.

Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T. S., & Yan, S. (2010). Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, *98*, 1031–1044.

Wu, L., Jin, R., Hoi, S.C., Zhu, J., & Yu, N. (2009). Learning Bregman distance functions and its application for semi-supervised clustering. In *Advances in neural information processing systems* (pp. 2089–2097).

Xu, L., & Yuille, A. L. (1995). Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Transactions on Neural Networks*, *6*, 131–143.

Yang, L., & Jin, R. (2006). *Distance metric learning: A comprehensive survey: 2*. Michigan State Universiy.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*, 301–320.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*, 265–286.