Robust Sparse Embedding and Reconstruction via Dictionary Learning

Konstantinos Slavakis and Georgios B. Giannakis
Digital Technology Center, University of Minnesota, USA
Emails: slavakis@dtc.umn.edu, georgios@umn.edu

Geert Leus
Delft University of Technology, The Netherlands
Email: g.j.t.leus@tudelft.nl

***Abstract*—A novel approach is developed for nonlinear compression and reconstruction of high- or even infinite-dimensional signals living on a smooth but otherwise unknown manifold. Compression is effected through affine embeddings to lower-dimensional spaces. These embeddings are obtained via linear regression and bilinear dictionary learning algorithms that leverage manifold smoothness as well as sparsity of the affine model and its residuals. The emergent unifying framework is general enough to encompass known locally linear embedding and compressive sampling approaches to dimensionality reduction. Emphasis is placed on reconstructing high-dimensional data from their low-dimensional embeddings. Preliminary tests demonstrate the analytical claims, and their potential to (de)compressing synthetic and real data.**

I. Introduction

Consider $N \in \mathbb{N}_*$ high-dimensional (column) vectors $\{\boldsymbol{x}_n\}_{n=1}^N \subset \mathbb{R}^D$, located on or close to a smooth but otherwise unknown manifold $\mathcal{M} \subset \mathbb{R}^D$, $D \in \mathbb{N}_*$. Given these training data vectors, critical for efficient source encoding and decoding of out-of-sample vectors $\boldsymbol{x}$ are: (a) the dimensionality reduction module, which effects (generally lossy) compression from high-dimensional ($\boldsymbol{x} \in \mathbb{R}^D$) to low-dimensional ($\boldsymbol{y} \in \mathbb{R}^d$, $\mathbb{N}_* \ni d \ll D$) vectors at the transmitter (Tx); as well as (b) the reconstruction module at the receiver (Rx), which yields estimates $\hat{\boldsymbol{x}}$ of the high-dimensional vectors from their low-dimensional renditions.

Principal component analysis (PCA) relies on the Karhunen-Loeve transform, which constitutes the "workhorse" of dimensionality reduction using a *linear* operator, namely a $d \times D$ matrix $\boldsymbol{V}^\top$ ($^\top$ denotes transposition) formed by the eigenvectors corresponding to the d (out of D) largest eigenvalues of the sample covariance matrix $N^{-1}\sum_{n=1}^N \boldsymbol{x}_n\boldsymbol{x}_n^\top$ [1, Chap. 14.5]. PCA's premise for compressing $\boldsymbol{x}$ to its lower-dimensional rendition $\boldsymbol{y} = \boldsymbol{V}^\top\boldsymbol{x}$ at the Tx, and reconstructing it optimally, in the mean-square sense, is that $\boldsymbol{x}$ is stationary with the same covariance matrix as $\{\boldsymbol{x}_n\}_{n=1}^N$. From a deterministic viewpoint, PCA is effective in (de)compression provided that both training and out-of-sample vectors live on (or stay close in the least-squares (LS) sense to) an affine subspace.

Data vectors, however, do not generally lie on an affine subspace but often on a manifold. In addition, they are typically realizations of nonstationary or locally stationary processes, including those formed by e.g., image and speech signals. These considerations motivate approaches to *nonlinear* dimensionality reduction [1, Chap. 14.9]. Among those, the so-termed locally linear embedding (LLE) approach has well-documented merits, because [2]: (a) it is computationally affordable, entailing closed-form expressions and eigen-decomposition level complexity; (b) it does not require knowledge but only smoothness of the manifold; and (c) it leverages smoothness to learn the manifold, and obtain LLEs that can be thought of as being applied on tangential affine subspaces.

So far, LLE has been advocated for manifold learning, clustering, and classification [3]. Recently, sparsity has been exploited for LLE-type robust manifold learning and low-dimensional embedding [4]–[7], but not for reconstruction purposes in a source (de)coding setup. Sparsity is also the enabling attribute for compressive sampling (CS) via random projections [8]–[13], dictionary learning (DL) [14], [15], and reconstruction, but its role for LLE-like nonlinear (de)compression has not been investigated.

The present paper aims to fill in these gaps, through a two-pronged objective: (a) Based on $\{\boldsymbol{x}_n\}_{n=1}^N$, the goal is to develop a sparsity-aware, outlier-resilient estimate of the manifold $\mathcal{M}$, and map it to a lower-dimensional space $\mathbb{R}^d$ (with $d \ll D$) by *robust sparse embeddings* (RSE) in the training phase; and (b) leverage this mapping during the operational phase to "compress" $\boldsymbol{x} \in \mathbb{R}^D$ as $\boldsymbol{y} \in \mathbb{R}^d$ at the Tx, and use the latter (or its noisy version $\hat{\boldsymbol{y}}$) to reconstruct an estimate $\hat{\boldsymbol{x}}$ of $\boldsymbol{x}$ at the Rx.

The rest of the paper is organized as follows. The basic principles of RSE are given in Section II, and its special case, the robust sparse global embedding (RSGE) is established in Section III. The more general approach of robust sparse embedding via dictionary learning (RSE-DL) is presented in Section IV. The RSE framework is further broadened in Section V to accommodate compression and reconstruction of continuous-time or continuous argument signals lying close to smooth manifolds in generally infinite dimensional Hilbert spaces. Numerical tests on synthetic and real data corroborate the analytical claims in Section VI, and conclusions are offered in Section VII.

For notational brevity, collect $\{\boldsymbol{x}_n\}_{n=1}^N$ to form the $D \times N$ matrix $\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$. The classical vector Euclidean norm is denoted by $\|\cdot\|$, while the ℓ_1-norm by $\|\cdot\|_1$. Moreover, given any integers j_1, j_2, with $j_1 \leq j_2$, let $\overline{j_1, j_2} := \{j_1, j_1+1, \ldots, j_2\}$.

II. Robust Sparse Embedding

Available to both Tx and Rx are the training data $\boldsymbol{X}$, and prescribed are the dimensions (D, d), with $d \ll D$, and the cardinality $K(\ll N)$, of the set of training data that can be approximately considered to lie on an affine subspace.

The first step during the *training phase* is to determine per datum $\boldsymbol{x}_n$ an $N \times 1$ weight vector $\boldsymbol{w}_n^x$ (with K_n nonzero entries $w_{n\nu}^x$), and a $D \times 1$ sparse bias vector $\boldsymbol{b}_n^x$. Based on $\boldsymbol{w}_n^x$ and $\boldsymbol{b}_n^x$,

This work was supported by the MURI Grant AFOSR FA9550-10-1-0567.

978-1-4673-5239-0/13/$31.00 ©2013 IEEE

each training datum can be approximated as an affine combination of its "affiliates," that is $\boldsymbol{x}_n \approx \sum_{\nu=1}^{N} w_{n\nu}^x \boldsymbol{x}_\nu + \boldsymbol{b}_n^x$. Vectors $(\boldsymbol{w}_n^x, \boldsymbol{b}_n^x)$ will be obtained by solving the least-absolute shrinkage and selection operator (Lasso)-type minimization task:

$$(\boldsymbol{w}_n^x, \boldsymbol{b}_n^x) = \arg \min_{\substack{(\boldsymbol{w},\boldsymbol{b})\in\mathbb{R}^N\times\mathbb{R}^D \\ \boldsymbol{w}^\top \mathbf{1}_N=1,\ w_n=0}} \left\| \boldsymbol{x}_n - \sum_{\nu=1}^{N} w_\nu \boldsymbol{x}_\nu - \boldsymbol{b} \right\|^2 + \lambda_{wn}\|\boldsymbol{w}\|_1 + \lambda_{bn}\|\boldsymbol{b}\|_1 \quad (1)$$

where $\mathbf{1}_N$ denotes the vector of all ones; $\lambda_{wn} \geq 0$ can be tuned to yield up to K_n nonzero entries in $\boldsymbol{w}_n^x$, and $\lambda_{bn} \geq 0$ controls the number of dimensions that can be offset to better fit the basis affiliates of $\boldsymbol{x}_n$ to an affine subspace tangential to the smooth $\mathcal{M}$ at $\boldsymbol{x}_n$. It is worth stressing that these scalars are generally n-dependent thus, e.g., allowing bases of variable size K_n per datum.

RSE bears similarities with LLE, which are instructive to recap in order to gain intuition. Suppose that $\{\boldsymbol{x}_n\}_{n=1}^N$ form clusters on $\mathcal{M}$, and say $\boldsymbol{x}_n$ belongs to one of these clusters. Thanks to smoothness, this cluster comprises points that can be thought of as residing on an affine subspace. Hence, $\boldsymbol{x}_n$ can be well approximated by an affine combination of other points in its cluster. As with LLE, the rationale behind RSE is that data (perhaps locally) belonging to an affine subspace can afford PCA-like linear dimensionality reduction. However, compressing matrices must be n-dependent and cannot be estimated reliably as in PCA since the number of points per affine subspace maybe too small.

Despite their similarities, there are also distinct differences between RSE and LLE. To start, RSE does not necessarily find K_n neighbors of each training datum based on their Euclidean distances from $\boldsymbol{x}_n$, but instead it lets the ℓ_1-norm of $\boldsymbol{w}_n$ select the K_n affiliates defining its affine subspace. This sparsity-promoting regularization has been also recognized by [4]–[7]. However, the auxiliary sparse vector $\boldsymbol{b}_n^x$ is unique to the RSE here and can be viewed as a "de-biasing" or outlier-capturing variable, as in the robust PCA and multi-dimensional scaling approaches of [16]. Correspondingly, regularization with the ℓ_1-norm of $\boldsymbol{b}_n^x$ renders the RSE- as well as the LLE-based manifold learning approaches, robust to outliers.

Instead of employing $\boldsymbol{b}_n^x$ and the LS cost in (1), robustness to outliers can be effected by adopting alternative costs such as the ϵ-insensitive one ($\epsilon > 0$), which is the criterion of choice for support vector regression. In this case, the following convex problem is solved per $\boldsymbol{x}_n$ [cf. (1)]

$$\min_{\substack{\boldsymbol{w}\in\mathbb{R}^N \\ \boldsymbol{w}^\top \mathbf{1}_N=1,\ w_n=0}} \sum_{\delta=1}^{D} \max\left\{0, \left| x_{n\delta} - \sum_{\nu=1}^{N} w_\nu x_{\nu\delta} \right| - \epsilon \right\} + \lambda_{wn}\|\boldsymbol{w}\|_1.$$

Having learned the manifold through its weights $\{\boldsymbol{w}_n^x\}_{n=1}^N$, RSE proceeds to find the nonlinear mapping which embeds $\mathcal{M}$ into $\mathbb{R}^d$ by solving the following constrained optimization problem:

$$\min_{\substack{\boldsymbol{Y}\in\mathbb{R}^{d\times N} \\ \boldsymbol{Y}\boldsymbol{Y}^\top=\boldsymbol{I}_d,\ \boldsymbol{Y}\mathbf{1}_N=\mathbf{0}_d}} \sum_{n=1}^{N} \left\| \boldsymbol{y}_n - \sum_{\nu=1}^{N} w_{n\nu}^x \boldsymbol{y}_\nu \right\|^2 \quad (2)$$

where $\boldsymbol{Y} := [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N]$; vector $\mathbf{0}_d$ is the $d \times 1$ all-zero vector; and $\boldsymbol{I}_d$ denotes the $d \times d$ identity matrix. The first constraint in (2) excludes the trivial solution $\boldsymbol{Y} = \mathbf{0}_{d\times N}$, while the second one centers the columns of $\boldsymbol{Y}$ around $\mathbf{0}_d$, since the cost in (2) is invariant to translation of the $\boldsymbol{Y}$ columns [cf. (1)]. Note also that if $\boldsymbol{Y}$ solves (2), then so does $\boldsymbol{U}\boldsymbol{Y}$ for any orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{d\times d}$. As in LLE, the solution of (2) is given by the eigen-decomposition of an appropriate $N \times N$ matrix [2]. Intuitively, this second step of the training phase ensures a globally consistent alignment of the (possibly local) affine subspace models of the tangential manifold patches.

Moving on to the *operational phase*, consider the out-of-sample vector $\boldsymbol{x}$. The first step is to learn the affine subspace $\boldsymbol{x}$ belongs to by solving the robust Lasso-type problem [cf. (1)]

$$(\boldsymbol{w}^x, \boldsymbol{b}^x) = \arg \min_{\substack{(\boldsymbol{w},\boldsymbol{b})\in\mathbb{R}^N\times\mathbb{R}^D \\ \boldsymbol{w}^\top \mathbf{1}_N=1}} \left\| \boldsymbol{x} - \sum_{n=1}^{N} w_n \boldsymbol{x}_n - \boldsymbol{b} \right\|^2 + \lambda_w\|\boldsymbol{w}\|_1 + \lambda_b\|\boldsymbol{b}\|_1 \quad (3)$$

where $\lambda_w, \lambda_b \geq 0$. Using this weight vector, the $d\times 1$ compressed version of $\boldsymbol{x}$ is readily obtained as

$$\boldsymbol{y} = \sum_{n=1}^{N} w_n^x \boldsymbol{y}_n = \boldsymbol{Y}\boldsymbol{w}^x \quad (4)$$

where $\boldsymbol{Y}$ is available from the training phase. This linear means of obtaining $\boldsymbol{y}$ replaces the computationally complex eigen-decomposition approach in (2).

At the Rx end, $\boldsymbol{y}$ may be received in noise as $\hat{\boldsymbol{y}}$. Nonetheless, it is possible to identify its affiliates in the d-space, by once again solving a Lasso-type problem:

$$(\hat{\boldsymbol{w}}^y, \hat{\boldsymbol{b}}^y) = \arg \min_{\substack{(\boldsymbol{w},\boldsymbol{b})\in\mathbb{R}^N\times\mathbb{R}^d \\ \boldsymbol{w}^\top \mathbf{1}_N=1}} \left\| \hat{\boldsymbol{y}} - \sum_{n=1}^{N} w_n \boldsymbol{y}_n - \boldsymbol{b} \right\|^2 + \lambda_w\|\boldsymbol{w}\|_1 + \lambda_b\|\boldsymbol{b}\|_1. \quad (5)$$

Upon recalling that the RSE weights in d-space remain invariant in D-space, it follows readily that $\boldsymbol{x}$ can be reconstructed as $\hat{\boldsymbol{x}} = \sum_{n=1}^{N} \hat{w}_n^y \boldsymbol{x}_n = \boldsymbol{X}\hat{\boldsymbol{w}}^y$.

III. RSGE

By construction, RSE embeds $\mathcal{M}$ in $\mathbb{R}^d$ so that affine relations among training vectors in $\mathbb{R}^D$ are preserved in $\mathbb{R}^d$ [cf. (1) and (2)]. RSE implicitly assumes the existence of a linear mapping $\boldsymbol{V}_n^\top$ per datum $\boldsymbol{x}_n$ (cf. the discussion on PCA in Section I), with weights $\boldsymbol{w}_n^x$ capturing affine relations in the cluster of $\boldsymbol{x}_n$ that are preserved after the embedding into $\mathbb{R}^d$. In contrast to the mappings $\{\boldsymbol{V}_n^\top\}_{n=1}^N$, each corresponding to a specific $\boldsymbol{x}_n$, the key idea of this section is to specialize RSE through a *global* linear mapping $\boldsymbol{A}^\top \in \mathbb{R}^{d\times D}$ that *explicitly* maps $\mathcal{M}$ to $\boldsymbol{A}^\top(\mathcal{M}) \subset \mathbb{R}^d$. Due to linearity, the affine relations of $\mathcal{M} \subset \mathbb{R}^D$ are also preserved in the embedded $\boldsymbol{A}^\top(\mathcal{M}) \subset \mathbb{R}^d$.

Care must be taken for $\boldsymbol{A}^\top$ not to deform the structure of $\mathcal{M}$ in $\mathbb{R}^d$. To this end, *stable* embedding operators $\boldsymbol{A}^\top$ have been developed in the context of compressive sampling (CS) to preserve, up to a scale factor, the original Euclidean distances of $\{\boldsymbol{x}_n\}_{n=1}^N$ [8]–[11]. Conditions on d, N, and $\boldsymbol{A}$ are reported in

[10], [11] to ensure that $\mathcal{M}$ is stably embedded in $\mathbb{R}^d$ with *high probability*.

Selecting such a stable embedding $\boldsymbol{A}^\top$, the compression operator of Section II can be replaced by $\boldsymbol{y}_n = \boldsymbol{A}^\top \boldsymbol{x}_n$, $\forall n \in \overline{1, N}$. As in the operational phase of Section II, for each out-of-sample $\boldsymbol{x} \in \mathbb{R}^D$, an optimization identical to (3) is solved at the Tx to find a sparse $\boldsymbol{w}^x$, based on which the compressed vector is obtained as $\boldsymbol{y} = \boldsymbol{A}^\top \left(\sum_{n=1}^N w_n^x \boldsymbol{x}_n \right) = \sum_{n=1}^N w_n^x \boldsymbol{y}_n = \boldsymbol{Y} \boldsymbol{w}^x$.

Vector $\boldsymbol{y}$ or its noisy version $\hat{\boldsymbol{y}}$ is available at the Rx, where a task identical to (5) is carried out. Since RSGE is a special case of RSE, and upon recalling that RSE weights in $\mathbb{R}^d$ coincide with those in $\mathbb{R}^D$, the reconstruction step yields $\hat{\boldsymbol{x}} = \sum_{n=1}^N \hat{w}_n^y \boldsymbol{x}_n = \boldsymbol{X} \hat{\boldsymbol{w}}^y$.

IV. RSE Based on Dictionary Learning

In certain applications it is conceivable that $\{\boldsymbol{x}_n\}_{n=1}^N$ do not contain only manifold-related information, but also noise and outliers. As a result, $\{\boldsymbol{x}_n\}_{n=1}^N$ are not located onto but rather "close" to $\mathcal{M}$. Moreover, if N is small or the data are non-uniformly distributed, $\{\boldsymbol{x}_n\}_{n=1}^N$ may not describe well the manifold $\mathcal{M}$. It is thus prudent to remove noise and outliers from $\{\boldsymbol{x}_n\}_{n=1}^N$, while leveraging the smoothness of $\mathcal{M}$ and affine relations among data subsets. Such an objective was pursued in Section II through $\boldsymbol{b}_n^x$.

In this section, a more general approach is followed by allowing each neighborhood to be described by a basis other than the one provided by the training data themselves. Specifically, a general dictionary $\boldsymbol{\mathcal{X}} := [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_Q] \in \mathbb{R}^{D \times Q}$ is sought, for some positive integer $Q < N$, such that the basis vectors $\{\boldsymbol{\chi}_q\}_{q=1}^Q$ for the entire manifold are located and distributed close to $\mathcal{M}$ in a more flexible manner than $\{\boldsymbol{x}_n\}_{n=1}^N$, while at the same time describing accurately $\{\boldsymbol{x}_n\}_{n=1}^N$ through an affine model. This dictionary learning (DL)-based RSE approach solves

$$\begin{aligned} \min_{\substack{\boldsymbol{\mathcal{X}} \in \mathbb{R}^{D \times Q},\ \boldsymbol{\Gamma} \in \mathbb{R}^{D \times N},\ \boldsymbol{B} \in \mathbb{R}^{D \times Q} \\ \boldsymbol{\Omega} \in \mathbb{R}^{Q \times N},\ \boldsymbol{\Omega}^\top \mathbf{1}_Q = \mathbf{1}_N \\ \boldsymbol{W} \in \mathbb{R}^{Q \times Q},\ \boldsymbol{W}^\top \mathbf{1}_Q = \mathbf{1}_Q \\ \mathrm{Diag}(\boldsymbol{W}) = \mathbf{0}}} & \sum_{n=1}^N \left\| \boldsymbol{x}_n - \sum_{q=1}^Q \omega_{nq} \boldsymbol{\chi}_q - \boldsymbol{\gamma}_n \right\|^2 \\ & + \lambda_{\mathcal{X}} \sum_{q=1}^Q \left\| \boldsymbol{\chi}_q - \sum_{\tau=1}^Q w_{q\tau} \boldsymbol{\chi}_\tau - \boldsymbol{b}_q \right\|^2 \\ & + \sum_{n=1}^N \left(\lambda_{\omega n} \|\boldsymbol{\omega}_n\|_1 + \lambda_{\gamma n} \|\boldsymbol{\gamma}_n\|_1 \right) \\ & + \sum_{q=1}^Q \left(\lambda_{wq} \|\boldsymbol{w}_q\|_1 + \lambda_{bq} \|\boldsymbol{b}_q\|_1 \right) \end{aligned} \tag{6}$$

where $\lambda_{\mathcal{X}} \geq 0$, $\boldsymbol{\Omega} := [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N] \in \mathbb{R}^{Q \times N}$, $\boldsymbol{\Gamma} := [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_N] \in \mathbb{R}^{D \times N}$, $\boldsymbol{W} := [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_Q] \in \mathbb{R}^{Q \times Q}$, $\boldsymbol{B} := [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_Q] \in \mathbb{R}^{D \times Q}$, and $\{\lambda_{\omega n}, \lambda_{\gamma n}\}_{n=1}^N$, $\{\lambda_{wq}, \lambda_{bq}\}_{q=1}^Q$ are non-negative real-valued parameters which control the sparsity level of the vectors in (6). Different from e.g., [15], RSE-DL in (6) is constrained by the smoothness of $\mathcal{M}$.

Once the sought dictionary $\boldsymbol{\mathcal{X}}$ becomes available from (6), embedding of $\mathcal{M}$ into $\mathbb{R}^d$ is possible by solving

$$\min_{\substack{\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{d \times Q} \\ \boldsymbol{\mathcal{Y}} \boldsymbol{\mathcal{Y}}^\top = \boldsymbol{I}_d,\ \boldsymbol{\mathcal{Y}} \mathbf{1}_Q = \mathbf{0}_d}} \sum_{q=1}^Q \left\| \boldsymbol{\psi}_q - \sum_{\tau=1}^Q w_{q\tau} \boldsymbol{\psi}_\tau \right\|^2 \tag{7}$$

where $\boldsymbol{\mathcal{Y}} := [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_Q] \in \mathbb{R}^{d \times Q}$. Upon defining $\boldsymbol{Y} := [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N] := \boldsymbol{\mathcal{Y}} \boldsymbol{\Omega}$, and having available $(\boldsymbol{X}, \boldsymbol{Y})$ at both Tx and Rx, the operational phases follow identical steps to those in Section II.

However, one can consider a reduced-complexity route (recall $Q < N$) by utilizing $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$ instead of $(\boldsymbol{X}, \boldsymbol{Y})$ as follows. In the operational phase, a procedure similar to (3) can be adopted, but with $\boldsymbol{\mathcal{X}}$ instead of $\boldsymbol{X}$; that is

$$\begin{aligned} (\boldsymbol{w}^x, \boldsymbol{b}^x) = \arg \min_{\substack{(\boldsymbol{w}, \boldsymbol{b}) \in \mathbb{R}^Q \times \mathbb{R}^D \\ \boldsymbol{w}^\top \mathbf{1}_Q = 1}} & \left\| \boldsymbol{x} - \sum_{q=1}^Q w_q \boldsymbol{\chi}_q - \boldsymbol{b} \right\|^2 \\ & + \lambda_w \|\boldsymbol{w}\|_1 + \lambda_b \|\boldsymbol{b}\|_1. \end{aligned} \tag{8}$$

Having available the dictionary $\boldsymbol{\mathcal{Y}}$ in the low-dimensional space $\mathbb{R}^d$, compression of $\boldsymbol{x}$ is achieved via $\boldsymbol{\psi} = \boldsymbol{\mathcal{Y}} \boldsymbol{w}^x$. At the Rx end, the potentially noisy $\hat{\boldsymbol{\psi}} \in \mathbb{R}^d$ is utilized to solve the following Lasso-type problem

$$\begin{aligned} (\hat{\boldsymbol{w}}^\psi, \hat{\boldsymbol{b}}^\psi) = \arg \min_{\substack{(\boldsymbol{w}, \boldsymbol{b}) \in \mathbb{R}^Q \times \mathbb{R}^D \\ \boldsymbol{w}^\top \mathbf{1}_Q = 1}} & \left\| \hat{\boldsymbol{\psi}} - \sum_{q=1}^Q w_q \boldsymbol{\psi}_q - \boldsymbol{b} \right\|^2 \\ & + \lambda_w \|\boldsymbol{w}\|_1 + \lambda_b \|\boldsymbol{b}\|_1. \end{aligned} \tag{9}$$

Finally, the original $\boldsymbol{x}$ is reconstructed at the Rx as $\hat{\boldsymbol{x}} := \sum_{q=1}^Q \hat{w}_q^\psi \boldsymbol{\chi}_q = \boldsymbol{\mathcal{X}} \hat{\boldsymbol{w}}^y$. Both the training and operational phases of RSE-DL are summarized in Fig. 1.

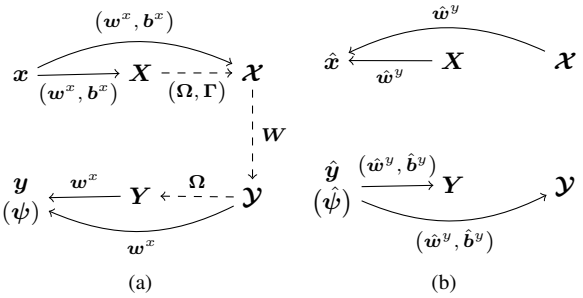

Fig. 1. (a) Training and operational phase of RSE-DL at Tx. Dashed lines correspond to the training phase, while solid lines to the operational one. Arrows pointing to the right precede in time. Moreover, bent lines stand for the optional less computationally demanding route discussed in Section IV. (b) Operational phase of RSE-DL at Rx. Bent lines stand for the optional less computational demanding route employing $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$.

Remark 1. If the solver of (8) yields vectors $\boldsymbol{w}^x$ with support having cardinality upper bounded by d, then $\boldsymbol{w}^x$ can be taken as the compressed version of $\boldsymbol{x}$ instead of $\boldsymbol{\psi}$. This resembles the DL-based sparse coding scheme of e.g., [15], and bypasses the need for solving (9). Vector $\boldsymbol{x}$ can be reconstructed as $\hat{\boldsymbol{x}} = \boldsymbol{\mathcal{X}} \hat{\boldsymbol{w}}^x$, where $\hat{\boldsymbol{w}}^x$ is the noisy version of the $\boldsymbol{w}^x$ sent to Rx. Of course, in addition to the nonzero entries of $\boldsymbol{w}^x$, their locations must be communicated from the Tx to the Rx.

Remark 2. The RSE-DL approach can be viewed as a deterministic alternative to the Bayesian multi-factor analysis (B-MFA) scheme in [12]; see also [13]. B-MFA utilizes training data to learn low-dimensional signal models (specifically to estimate (hyper)parameters of Gaussian mixture priors based on costly Monte Carlo sampling). But unlike RSE-DL which relies on data-adaptive training to find the embedding, B-MFA relies on *data non-adaptive* CS measurement matrices to obtain the embedding. As such, B-MFA is neither tailored for clustering and classification based on low-dimensional features, nor for (de)compression especially when the reduced dimension renders the CS matrix information lossy.

V. RSE in Hilbert Spaces

The scope of RSE is broadened in this section to dimensionality reduction and reconstruction of signals belonging even to infinite dimensional (real) Hilbert spaces $\mathcal{H}$ instead of $\mathbb{R}^D$. Rather than vectors $\boldsymbol{x} \in \mathbb{R}^D$, of interest here are compression and reconstruction of (nonlinear) functions $\mathbb{R}^p \ni \boldsymbol{t} \mapsto f(\boldsymbol{t}) \in \mathbb{R}$, for some $p \in \mathbb{N}_*$, which are located onto or close to a smooth but unknown manifold $\mathcal{M} \subset \mathcal{H}$. Space $\mathcal{H}$ is assumed equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and induced norm $\|\cdot\|_{\mathcal{H}}$.

In this context, consider a training set of functions $\{f_n\}_{n=1}^N \subset \mathcal{H}$. Parallel to vector $\boldsymbol{b}^x$ in Section II, $M \in \mathbb{N}_*$ user-defined functions $\{g_m\}_{m=1}^M \subset \mathcal{H}$ are also considered to approximate the residual $f - \sum_{\nu=1}^N w_\nu^f f_\nu$. With $\{w_\nu^f\}_{\nu=1}^N$ and $\{v_m^f\}_{m=1}^M$ denoting scalar weights, $f \in \mathcal{H}$ located close to $\mathcal{M}$ is expressed as

$$f \approx \sum_{\nu=1}^N w_\nu^f f_\nu + \sum_{m=1}^M v_m^f g_m \ . \tag{10}$$

It can be readily verified that application of *any* linear mapping $S : \mathcal{H} \to \mathbb{R}^D : f \mapsto S(f)$ onto (10) yields $S(f) \approx \sum_{\nu=1}^N w_\nu^f S(f_\nu) + \sum_{m=1}^M v_m^f S(g_m)$. In other words, affine relations in $\mathcal{M} \subset \mathcal{H}$ are preserved in $S(\mathcal{M}) \subset \mathbb{R}^D$. For specificity, consider the following example.

Example 1. The mapping $\mathcal{H} \ni f \mapsto S_{\mathcal{T}}(f) := [f(\boldsymbol{t}_1), \ldots, f(\boldsymbol{t}_D)]^\top \in \mathbb{R}^D$ is linear, if e.g., $\mathcal{T} := \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_D\} \subset \mathbb{R}^p$ denotes a generally non-uniform set of sampling points.

Compression and reconstruction of functions in $\mathcal{H}$ is transformed via $S_{\mathcal{T}}$ into a corresponding problem in the finite dimensional $\mathbb{R}^D$. The original task is relaxed from everywhere approximation in (10) to an interpolation problem onto a finite number of sampling points $\mathcal{T}$. Upon defining $\boldsymbol{x}_n := S_{\mathcal{T}}(f_n)$, $\forall n \in \overline{1, N}$, the framework of Sections II and IV can be readily applied to the problem in $\mathbb{R}^D$. For example, (1) translates to

$$\begin{aligned}(\boldsymbol{w}_n^f, \boldsymbol{b}_n^f) = \arg \min_{(\boldsymbol{w}, \boldsymbol{b})} & \left\| S_{\mathcal{T}}(f_n) - \sum_{\nu=1}^N w_\nu S_{\mathcal{T}}(f_\nu) - \boldsymbol{b} \right\|^2 \\ & + \lambda_{wn} \|\boldsymbol{w}\|_1 + \lambda_{bn} \|\boldsymbol{b}\|_1 \\ \text{s.t.} & \begin{cases} (\boldsymbol{w}, \boldsymbol{b}) \in \mathbb{R}^N \times \mathbb{R}^D, \\ \boldsymbol{w}^\top \mathbf{1}_N = 1, \ w_n = 0 \end{cases}\end{aligned}$$

where $\boldsymbol{b}$ was used to generalize $\sum_{m=1}^M v_m^f S_{\mathcal{T}}(g_m)$.

Since the only requirement on S is linearity, there are more ways to specify S other than that of Example 1.

Example 2. Given $\Phi := \{\varphi_1, \ldots, \varphi_D\} \subset \mathcal{H}$, define the linear operator $\mathcal{H} \ni f \mapsto S_\Phi(f) := [\langle f, \varphi_1 \rangle_{\mathcal{H}}, \ldots, \langle f, \varphi_D \rangle_{\mathcal{H}}]^\top \in \mathbb{R}^D$. Here too the results of Sections II and IV can be directly applied to the problem at hand after defining $\boldsymbol{x}_n := S_\Phi(f_n)$, $\forall n \in \overline{1, N}$.

There is a special class of $\mathcal{H}$ where the inner products $\{\langle f, \varphi_\delta \rangle\}_{\delta=1}^D$ in Example 2 are easily obtained. Indeed, if $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) [17], [18], $S_{\mathcal{T}}$ and S_Φ can be made to coincide. Recall that $\mathcal{H}$ is RKHS iff there exists a (unique) kernel function $\kappa(\cdot, \cdot) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ such that (i) $\kappa(\cdot, \boldsymbol{t}) \in \mathcal{H}$, and (ii) the *reproducing property* holds, i.e., $\langle f, \kappa(\cdot, \boldsymbol{t}) \rangle_{\mathcal{H}} = f(\boldsymbol{t})$, $\forall f \in \mathcal{H}$, $\forall \boldsymbol{t} \in \mathbb{R}^p$.

Example 3. Let $\mathcal{H}$ be RKHS, with kernel κ, and $\Phi := \{\kappa(\cdot, \boldsymbol{t}_1), \ldots, \kappa(\cdot, \boldsymbol{t}_D)\}$. Then, by the reproducing property, $S_\Phi(f) = [f(\boldsymbol{t}_1), \ldots, f(\boldsymbol{t}_D)] = S_{\mathcal{T}}(f)$, $\forall f \in \mathcal{H}$.

It is worth stressing here that S is no longer required and that RSE-type operations become viable in the original $\mathcal{H}$, if inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ are realizable. For example, following (1), the first step of the RSE's training phase in $\mathcal{H}$ is formulated as

$$\begin{aligned}(\boldsymbol{w}_n^f, \boldsymbol{v}_n^f) = \arg \min_{(\boldsymbol{w}^f, \boldsymbol{v}^f)} & \left\| f_n - \sum_{\nu=1}^N w_\nu^f f_\nu - \sum_{m=1}^M v_m^f g_m \right\|_{\mathcal{H}}^2 \\ & + \lambda_{wn} \|\boldsymbol{w}^f\|_1 + \lambda_{vn} \|\boldsymbol{v}^f\|_1 \\ \text{s.t.} & \begin{cases} (\boldsymbol{w}^f, \boldsymbol{v}^f) \in \mathbb{R}^N \times \mathbb{R}^M, \\ \mathbf{1}_N^\top \boldsymbol{w}^f = 1, \ w_n^f = 0. \end{cases}\end{aligned} \tag{11}$$

If $\langle f_{n_1}, f_{n_2} \rangle_{\mathcal{H}}$, $\forall (n_1, n_2) \in \overline{1, N}^2$, and $\langle f_n, g_m \rangle_{\mathcal{H}}$, $\forall (n, m) \in \overline{1, N} \times \overline{1, M}$ are assumed available, then (11) is equivalent to

$$\begin{aligned}(\boldsymbol{w}_n^f, \boldsymbol{v}_n^f) = \arg \min_{(\boldsymbol{w}^f, \boldsymbol{v}^f)} & \begin{bmatrix} \boldsymbol{w}^{f\top}, \boldsymbol{v}^{f\top} \end{bmatrix} \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{12}^\top & \boldsymbol{C}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}^f \\ \boldsymbol{v}^f \end{bmatrix} \\ & - 2 \begin{bmatrix} \boldsymbol{w}^{f\top}, \boldsymbol{v}^{f\top} \end{bmatrix} \begin{bmatrix} \boldsymbol{h}_1 \\ \boldsymbol{h}_2 \end{bmatrix} + \lambda_{wn} \|\boldsymbol{w}^f\|_1 + \lambda_{vn} \|\boldsymbol{v}^f\|_1 \\ \text{s.t.} & \begin{cases} (\boldsymbol{w}^f, \boldsymbol{v}^f) \in \mathbb{R}^N \times \mathbb{R}^M, \\ \mathbf{1}_N^\top \boldsymbol{w}^f = 1, \ w_n^f = 0 \end{cases}\end{aligned}$$

where $\boldsymbol{C}_{11} := [\langle f_i, f_j \rangle_{\mathcal{H}}]_{(i,j) \in \overline{1,N}^2} \in \mathbb{R}^{N \times N}$, $\boldsymbol{C}_{12} := [\langle f_i, g_j \rangle_{\mathcal{H}}]_{(i,j) \in \overline{1,N} \times \overline{1,M}} \in \mathbb{R}^{N \times M}$, $\boldsymbol{C}_{22} := [\langle g_i, g_j \rangle_{\mathcal{H}}]_{(i,j) \in \overline{1,M}^2} \in \mathbb{R}^{M \times M}$, and $\boldsymbol{h}_1 := [\langle f_n, f_i \rangle_{\mathcal{H}}]_{i \in \overline{1,N}} \in \mathbb{R}^N$, $\boldsymbol{h}_2 := [\langle f_n, g_i \rangle_{\mathcal{H}}]_{i \in \overline{1,M}} \in \mathbb{R}^M$. Similar derivations are also possible for the RSE-type operational phase in $\mathcal{H}$, provided that $\{\langle f, f_n \rangle_{\mathcal{H}}\}_{n=1}^N$ and $\{\langle f, g_m \rangle_{\mathcal{H}}\}_{m=1}^M$ are realizable for every $f \in \mathcal{H}$.

VI. Numerical Examples

To validate the proposed RSE family of algorithms, synthetic and real data are utilized from the celebrated "Swissroll manifold" and "Frey faces" database, respectively [19]. For comparison, PCA and LLE are also tested.

For the Swissroll data, $D = 3$ and $d = 2$. A number of $N = 500$ training and $N_{\text{test}} = 500$ out-of-sample data were generated by adding zero-mean Gaussian noise, at SNR $= 20$dB, on 3-dimensional samples taken randomly out of the manifold. Zero-mean Gaussian noise at SNR $= 10$dB was added to $\boldsymbol{y}$ to obtain $\hat{\boldsymbol{y}}$. The parameters used for the RSE family of algorithms are as follows: $\lambda_{wn} = 10^{-1}$, $\lambda_{bn} = 0$, $\lambda_{\omega n} = 10^{-1}$, $\lambda_{\gamma n} = 0$,

Method	MSE (dB)	Method	MSE (dB)
PCA	-10.395	RSE	-13.748
LLE	$+3.382$	RSE-DL $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$	-15.707
RSGE	-9.459	RSE-DL $(\boldsymbol{X}, \boldsymbol{Y})$	-13.766

TABLE I
MSEs ON 100 REALIZATIONS. THE "RSE-DL $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$" TAG REFERS TO THE DISCUSSION IN SECTION IV REGARDING THE REDUCED-COMPLEXITY PROMOTING DICTIONARIES $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$ OF CARDINALITY Q; SEE ALSO FIG. 1.

$\forall n \in \overline{1, N}$, while $\lambda_{\mathcal{X}} = 1$, $\boldsymbol{\Gamma} = \boldsymbol{B} = \boldsymbol{0}$, and $Q = 250$ for RSE-DL in Section IV. In LLE, $K = 20$ neighbors were used. Both embedding and reconstruction results are depicted in Fig. 2. To assess performance, the normalized reconstruction error was

$$\text{MSE} := \frac{1}{100 N_{\text{test}}} \sum_{r=1}^{100} \sum_{l=1}^{N_{\text{test}}} \frac{\|\boldsymbol{x}_l^{(r)} - \hat{\boldsymbol{x}}_l^{(r)}\|^2}{\|\boldsymbol{x}_l^{(r)}\|^2}$$

where r indexes the realization of the experiment. A number of 100 realizations were conducted to yield the results of Table I. Both in Fig. 2 and Table I, the tag "RSE-DL $(\boldsymbol{X}, \boldsymbol{Y})$" refers to Section IV where all N training data $\boldsymbol{X}$, and their "compressed" counterparts $\boldsymbol{Y}$ are utilized, while "RSE-DL $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$" indicates the reduced complexity approach of Section IV, where the Q-cardinality dictionaries $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$ are employed.

Results on embedding a number of $N = 1965$, (20×28)-dimensional ($D = 560$) "Frey faces" in $\mathbb{R}^2$ are shown in Fig. 3. In this context, the following values for the parameters in the RSE-family of algorithms are utilized: $\lambda_{wn} = 10^{-1}$, $\lambda_{bn} = 10^{-1}$, $\lambda_{\omega n} = 10^{-1}$, and $\lambda_{\gamma n} = 0$, $\forall n \in \overline{1, N}$, while $\lambda_{\mathcal{X}} = 1$, $\boldsymbol{\Gamma} = \boldsymbol{B} = \boldsymbol{0}$, and $Q = 1500$ for the RSE-DL approach of Section IV. Again, $K = 20$ is used for the LLE.

The numerical results of Figs. 2, 3 and Table I corroborate the analytical claims, and reveal the potential of RSE and RSE-DL for reconstruction, clustering and classification tasks.

VII. CONCLUSIONS

A robust, sparsity-aware, data-adaptive, nonlinear embedding methodology was developed for (de)compression of high- or even infinite-dimensional signals located close to smooth but otherwise unknown manifolds. The novel RSE approach enabled unsupervised learning of both local and global geometrical characteristics of smooth manifolds with computationally affordable training and operational phases. Moreover, the introduced RSE-DL alternative offered desirable morphing of the training data and noise mitigation to obtain a smoother basis for the manifold with reduced computational complexity. Numerical results were also presented to support the analytical claims.

Future research directions of prime interest include performance analysis results on the RSE and RSE-DL based reconstruction schemes, and thorough comparisons with existing alternatives on real images.

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[2] L. K. Saul and S. T. Roweis, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.

[3] A. Ghodsi, "Dimensionality reduction: A short tutorial," University of Waterloo, Tech. Rep. 2006-14, 2006.

[4] S. Huang, C. Cai, and Y. Zhang, "Dimensionality reduction by using sparse reconstruction embedding," in *Lecture Notes in Computer Science*, vol. 6298, 2010, pp. 167–178.

[5] R. Timofte and L. van Gool, "Sparse representation based projections," in *Proc. of the British Machine Vision Conf.*, 2011.

[6] E. Elhamifar and R. Vidal, "Sparse manifold clustering and embedding," in *Proc. of Neural Inf. Process. Syst.*, Spain, Dec. 2011.

[7] D. Kong, C. Ding, H. Huang, and F. Nie, "An iterative locally linear embedding algorithm," in *Proc. of the Int. Conf. on Machine Learning*, 2012, http://arxiv.org/abs/1206.6463.

[8] R. G. Baraniuk and M. B. Wakin, "Random projections of smooth manifolds," *Found. of Comput. Math.*, vol. 9, no. 1, pp. 51–77, Feb. 2009.

[9] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. and Machine Intelligence*, vol. 31, no. 2, pp. 1–18, Feb. 2009.

[10] H. L. Yap, M. B. Wakin, and C. J. Rozell, "Stable manifold embeddings with operators satisfying the restricted isometry property," in *Proc. of the 45th Annual Conf. on Inf. Sc. and Systems*, Baltimore: Maryland, Mar. 2011.

[11] ——, "Stable manifold embeddings with structured random matrices," http://arxiv.org/abs/1209.3312, 2012.

[12] M. Chen, J. Silva, J. Paisley, C. Wang, D. Dunson, and L. Carin, "Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 6140–6155, Dec. 2010.

[13] L. Carin, R. G. Baraniuk, V. Cevher, D. Dunson, M. I. Jordan, G. Sapiro, and M. B. Wakin, "Learning low-dimensional signal models," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 39–51, Mar. 2011.

[14] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

[15] I. Tošić and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, Mar. 2011.

[16] P. Forero and G. B. Giannakis, "Sparsity-exploiting robust multidimensional scaling," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4118–4134, Aug. 2012.

[17] K. Slavakis, P. Bouboulis, and S. Theodoridis, "Online learning in reproducing kernel Hilbert spaces," in *E-Ref. Signal Processing*. Elsevier, 2013, to appear.

[18] B. Schölkopf and A. J. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2001.

[19] http://www.cs.nyu.edu/~roweis/.

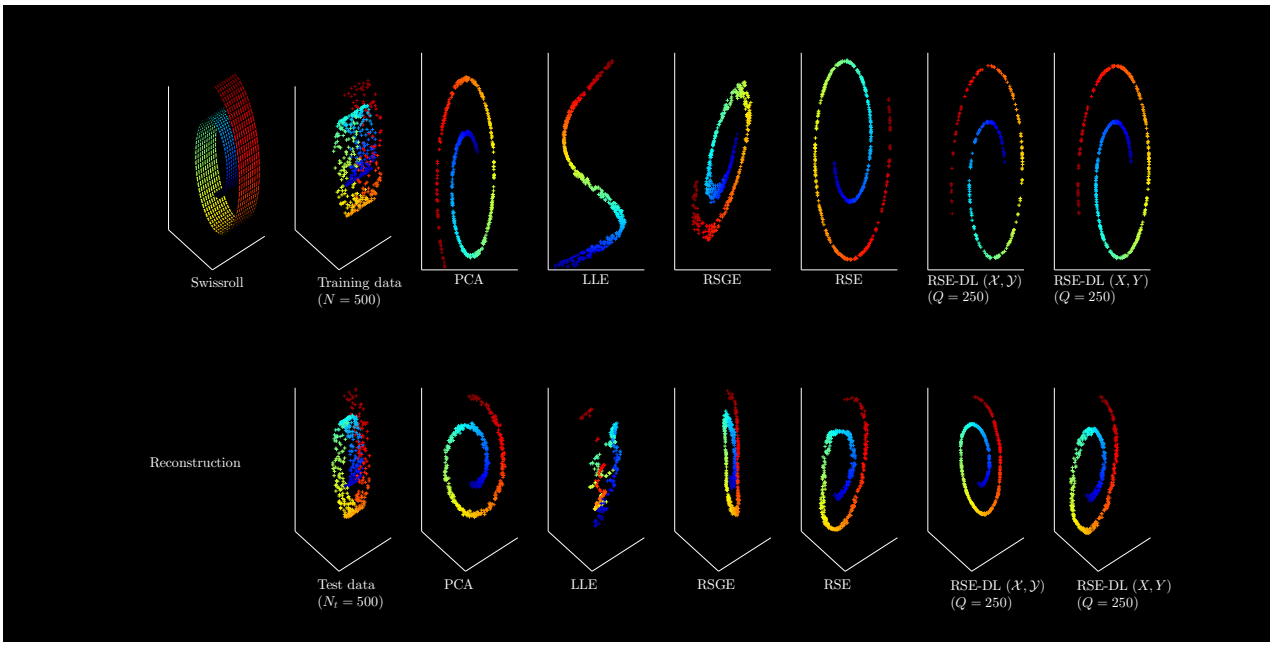

Fig. 2. Embedding of the Swissroll manifold in $\mathbb{R}^2$ and reconstruction.

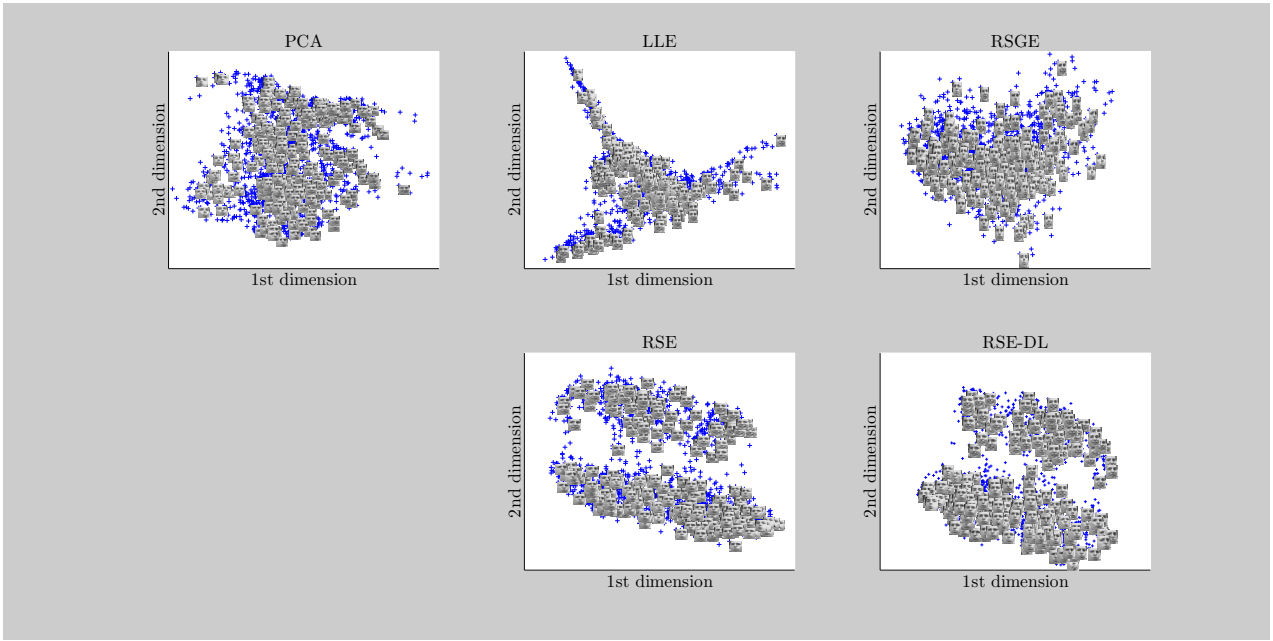

Fig. 3. Embedding of "Frey faces" in $\mathbb{R}^2$.