Multi-Resolutive Sparse Approximations of d-Dimensional Data

Giuseppe Patané

Consiglio Nazionale delle Ricerche
Istituto di Matematica Applicata e Tecnologie Informatiche
Genova, Italy

Abstract

This paper proposes an iterative computation of sparse representations of functions defined on $\mathbb{R}^d$, which exploits a formulation of the sparsification problem equivalent to Support Vector Machine and based on Tikhonov regularization. Through this equivalent formulation, the sparsification reduces to an approximation problem with a Tikhonov regularizer, which selects the null coefficients of the resulting approximation. The proposed multi-resolutive sparsification achieves a different resolution in the approximation of the input data through a hierarchy of nested approximation spaces. The idea behind our approach is to combine a smooth and strictly convex approximation of the l_1-norm with Tikhonov regularization and iterative solvers of linear/non-linear equations. Firstly, the iterative sparsification scheme is introduced in a Reproducing Kernel Hilbert Space with respect to its native norm. Then, the sparsification is generalized to arbitrary function spaces using the least-squares norm and radial basis functions. Finally, the discrete sparsification is derived using the eigendecomposition and the spectral properties of sparse matrices; in this case, the computational cost is $O(n \log n)$, with n number of input points. Assuming that the data is supported on a $(d-1)$-dimensional manifold, we derive a variant of the sparsification scheme that guarantees the smoothness of the solution in the ambient and intrinsic space by using spectral graph theory and manifold learning techniques. Finally, we discuss the multi-resolutive approximation of d-dimensional data such as signals, images, and 3D

Email address: patane@ge.imati.cnr.it (Giuseppe Patané)

shapes.

Keywords: Sparse approximation, Support Vector Machine, image analysis, least-squares approximation, Reproducing Kernel Hilbert Space, radial basis functions, spectral graph theory, manifold learning.

1. Introduction

Representing a signal as a linear combination of a set of atoms of a given dictionary is used in a wide range of applications, such as approximation, denoising, and compression. Two main elements characterize the final representation: (i) the properties of the atoms such as linear independence, orthogonality, redundancy, signal-awareness and (ii) the sparseness of the linear representation, which is given by the number of non-null coefficients. Defining sparse representations with respect to dictionaries reacher than an orthogonal basis is also fundamental to represent complex data and to adapt this representation to the features of the data itself. For instance, dictionaries of curvelets [8, 9] and bandelets [30, 46] are tailored to the local geometric regularity of the input signal and the coefficients of the corresponding sparse representations are useful to identify geometric features; e.g., sharp boundaries and edge orientation in images. Furthermore, the computation of sparse representations with respect to a given dictionary can be combined with an update of its atoms in order to improve the data fitting [1]. Main applications of sparse representations in computer vision and image understanding include face recognition [61], data segmentation [20, 50], image super-resolution [62], denoising [37], and classification [35, 36].

Given a signal $f:\mathbb{R}^d\rightarrow\mathbb{R}$ and a dictionary $\mathcal{B}:=\{\varphi_i(\mathbf{x})\}_{i=1}^n$ of atoms, sparse coding refers to the problem of computing the coefficients $\mathbf{a}:=(a_i)_{i=1}^n$ of the function $g(\mathbf{x})=\sum_{i=1}^n a_i\varphi_i(\mathbf{x})$ that approximates f, involves the smallest number of atoms, and provides the highest accuracy among all the approximations of f generated by $\mathcal{B}$. In this context, compressive sampling theory [8, 15] has shown that signals can be accurately approximated from a number of samples that is lower than the one imposed by the Nyquist sampling theory.

According to [14, 52], the coefficient vector $\mathbf{a}$, which defines the *sparse represen-*

tation $g : \mathbb{R}^d \to \mathbb{R}$ of f, solves the minimization problem

$$\arg\min_{\mathbf{a}\in\mathbb{R}^n} \{E(f,g) + \varepsilon\|\mathbf{a}\|_0\}, \qquad g(\mathbf{x}) := \sum_{i=1}^{n} a_i\varphi_i(\mathbf{x}), \tag{1}$$

where the term $E(f,g)$ is the approximation error between f and g with respect to the loss function $E(\cdot,\cdot)$; the sparsification order $\|\mathbf{a}\|_0$ is given by the number of non-null coefficients; and the positive constant ε controls the trade-off between these two terms.

To measure the approximation error between the maps f and g, common choices are the native distance $E(f,g) := \frac{1}{2}\|f-g\|_{\mathcal{H}}^2$ in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$; the l_2-norm $E(f,g) := \frac{1}{2}\|\mathbf{f}-\mathbf{g}\|_2^2$ of the values $\mathbf{f} := (f(\mathbf{x}_i))_{i=1}^n$, $\mathbf{g} := (g(\mathbf{x}_i))_{i=1}^n$ at the points of $\mathcal{P} := \{\mathbf{x}_i\}_{i=1}^n$; and the ε-insensitive cost function [14, 52]

$$E(f,g) := \sum_{i=1}^{n} \Gamma(f(\mathbf{x}_i) - g(\mathbf{x}_i)), \qquad \Gamma(t) := |t|_\varepsilon := \begin{cases} 0 & \text{if } |t| < \varepsilon, \\ |t| - \varepsilon & \text{otherwise.} \end{cases}$$

Since the minimization of the objective function in Eq. (1) is NP-hard, the sparsification term $\|\mathbf{a}\|_0$ is usually approximated by the l_p-norm $\|\mathbf{a}\|_p := \left(\sum_{i=1}^n |a_i|^p\right)^{1/p}$ and the corresponding sparsification results in a convex minimization problem. On the one hand, for $0 \leq p < 1$ the l_p-norm is not strictly convex [25, 47, 48, 49] and the corresponding problem has local extrema that might be identified as solutions during the search of the global minimum. On the other hand, the l_1-norm guarantees the uniqueness of the solution and provides a representation sparser than the l_2-norm. To avoid oversampling, an iterative re-weighted l_2-norm minimization, which provides a sparsity percentage lower than the l_2-norm, has been proposed in [28]. From a general perspective, the l_1-norm is preferable to the l_2-norm sparsification term because the former avoids to penalize outliers in the sampled data and to distribute the residual error in the objective functional [8, 15, 58]. Although the l_0-norm provides the sparsest solution, the assumption of dealing with a bounded noise generally guarantees that the l_1-norm sparse representations are significative and stable to noise and outliers.

The basis pursuit de-noising [11], regularized logistic regression [24, 41, 45, 51], standard [39] and orthogonality matching pursuit methods [10, 38, 43] use the l_1-norm as sparsification term. Since the l_1-norm is not differentiable at zero, the sparsification is converted to a constrained optimization problem [32], whose number of unknowns is twice the number of input variables. Alternatively, the l_1-norm is approximated by a second order Taylor expansion of the objective function [24], which is minimized using the least-squares angle regression [18] and the quasi-Newton algorithm [2]. The sparsification problem is also solved through an incremental approach [32], which is based on the conjugate gradient and avoids the discontinuity of the first order derivatives of the l_1-norm. Alternative approaches apply the maximum *a-posteriori* estimation [33, 34, 42] and uncertainty criteria [16, 17, 19, 21, 26]. Finally, the probabilistic Bayesian learning framework [57] is capable of further increasing the sparsification rate with respect to SVMs and applies to arbitrary kernels.

Aims and contributions. This paper discusses an iterative computation of sparse and multi-resolutive representations of an arbitrary function, which achieves a different resolution through a hierarchy of nested approximation spaces. The proposed approach exploits a formulation [22] of the sparse approximation problem equivalent to Support Vector Machine and based on Tikhonov regularization. Through this equivalent formulation, the sparsification reduces to an approximation problem with a Tikhonov regularizer, which selects the null coefficients of the resulting approximation. The idea behind our sparsification is to combine a smooth and strictly convex approximation of the l_1-norm with Tikhonov regularization and iterative solvers of linear or non-linear equations. The proposed approach also guarantees good generalization performances and applies to arbitrary function spaces, whose basis is not necessarily associated to a Mercer kernel. Finally, it provides a sequence of nested approximation spaces, which are generated by those functions selected during the computation of the sparsified solution. We also discuss the multi-resolutive approximation of d-dimensional data such as signals, images, and 3D shapes.

To introduce the sparsification scheme, we firstly assume that $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS) [3]; in this case, the native norm of $\mathcal{H}$ allows us to enforce the accuracy and smoothness of the sparse approximation. Using the equivalence between Support Vector Machine and Tikhonov regularization [22] in a RKHS, we approximate a real-valued function with sparse linear models, whose coefficients are fitted using a smoothed version of the l_1-regularization. This aim is achieved by replacing the l_1-norm with a smooth and strictly convex approximation; then, the corresponding sparsification functional is exactly evaluated and no approximation is required. Finally, the sparsification problem is converted into a system of non-linear equations, whose sparse coefficient vector is computed by applying a fixed point iteration and solving a sequence of linear systems.

Using radial basis functions and least-squares techniques, the second part of the paper generalizes the iterative sparsification scheme to arbitrary function spaces, which are not necessarily associated to Mercer kernels. Assuming that the data is supported on a $(d-1)$-dimensional manifold, we also derive a variant of the proposed approach that guarantees the smoothness of the solution in the ambient and intrinsic space by using spectral graph theory and manifold learning techniques. Diagonalizing the Gram matrix of the sparsification normal equation, the unknown coefficients become independent; i.e., each non-linear equation involves only one unknown and its solution is computed in explicit form.

Applying iterative solvers instead of decomposition methods for constrained convex minimization problems has the following advantages with respect to previous work. The computational cost of the overall framework is $O(r(n+n\log n))$ instead of $O(n^{3.5})$, where n and r, $r << n$, are the number of input data and steps of the iterative sparsification scheme, respectively. The solution of the sparsification system is well-conditioned as a matter of the underlying regularization framework and based on a global sparsification procedure, which avoids time-consuming and *a-posteriori* local updates of the model. Furthermore, at each iteration the update of the coefficient matrix involves only its diagonal elements, takes $O(n)$ time, and preserves its sparsity and symmetric structure. Finally, the input variables are not duplicated, thus reducing the memory allocation, which is one of the main drawbacks in case of a large amount of data. Since each iteration provides an approximate reconstruction of the input function $f:\mathbb{R}^d \rightarrow \mathbb{R}$, the iterative solver induces a hierarchy of sparse representations $(g^{(r)})_r$ of f, which belong to a sequence of nested spaces $(\mathcal{H}_r)_r$, $\mathcal{H}_{r+1} \subseteq \mathcal{H}_r$, $r \geq 1$.

The paper is organized as follows. First, we introduce the proposed sparsification scheme in Reproducing Kernel Hilbert Spaces (Sect. 2). Then, we derive a least-squares variant and its discrete counterpart (Sect. 3). Finally, we outline open issues and future work (Sect. 4).

2. “Iterative" sparse approximation in Reproducing Kernel Hilbert Spaces

Replacing the l_1-norm with a smooth approximation, we define an iterative sparsification scheme (Sect. 2.1) in a RKHS with respect to its native norm. Then, we discuss the iterative computation of the sparsified solution (Sect. 2.2), and the multi-resolutive structure of the sparsification scheme (Sect. 2.3). Finally, the generalization of the sparsification scheme to arbitrary function spaces is addressed in Section 3.

2.1. Sparsification in Reproducing Kernel Hilbert Spaces

Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space [3] endowed with the scalar product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$ induced by a positive definite, symmetric kernel $K:\mathbb{R}^d\times\mathbb{R}^d\rightarrow\mathbb{R}$. Common choices of $K(\cdot,\cdot)$ are the Gaussian $K(\mathbf{x},\mathbf{y}):=\exp(-\|\mathbf{x}-\mathbf{y}\|_2^2)$, polynomial $K(\mathbf{x},\mathbf{y}):=(1-\langle\mathbf{x},\mathbf{y}\rangle_2)^s$, and compactly supported [40, 53] kernels. Let $g:\mathbb{R}^d\rightarrow\mathbb{R}$

$$g(\mathbf{x}):=\sum_{i=1}^{n}a_iK(\mathbf{x},\mathbf{x}_i),\qquad \mathbf{a}:=(a_i)_{i=1}^n\in\mathbb{R}^n,$$

be a map in the linear space $\mathcal{H}_n\subseteq\mathcal{H}$ generated by the basis $\mathcal{B}:=\{\varphi_i(\mathbf{x})\}_{i=1}^n$, where each function $\varphi_i(\mathbf{x}):=K(\mathbf{x},\mathbf{x}_i)$ is induced by the kernel $K(\cdot,\cdot)$ and centered at the points of $\mathcal{P}:=\{\mathbf{x}_i\}_{i=1}^n$. The map g that provides the best compromise between approximation accuracy, which is measured by $E(f,g)=\frac{1}{2}\|f-g\|_{\mathcal{H}}$, and sparseness with respect to the l_1-norm is the solution to the minimization problem

$$\arg\min_{\mathbf{a}\in\mathbb{R}^n}\left\{\frac{1}{2}\left\|f-\sum_{i=1}^{n}a_iK(\mathbf{x},\mathbf{x}_i)\right\|_{\mathcal{H}}^2+\varepsilon\|\mathbf{a}\|_1\right\}.\qquad (2)$$

According to the equivalence between Support Vector Machine and Tikhonov regularization [22], the functional in Eq. (1) can be interpreted as a least-squares approximation problem along with the Tikhonov regularizer $\|\mathbf{a}\|_1$, which controls the sparsity of the corresponding solution. Since the function $s_\eta(t):=(t^2+\eta)^{1/2}$ approximates $|t|$ as $\eta\rightarrow0^+$, in $\mathbb{R}^n$ the l_1-norm $\|\mathbf{a}\|_1$ is approximated by $\psi(\mathbf{a}):=\sum_{i=1}^n(a_i^2+\eta)^{1/2}$, $\eta\rightarrow0^+$, (Fig. 1). Furthermore, from the upper bound $|\,|t|-s_\eta(t)\,|\leq\eta^{1/2}$, $\eta>0$,

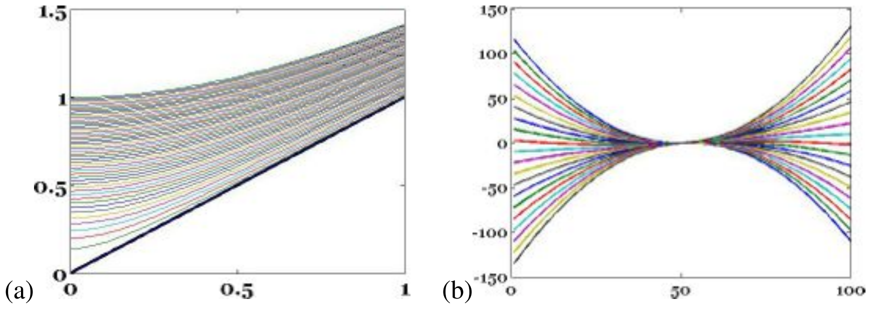


Figure 1: (a) Convergence of the function $s_\eta(t) := (t^2+\eta)^{1/2}$ to $|t|$, as $\eta \to 0^+$, $t \in [0,1]$. Each plot of s_η corresponds to a different value of the η. (b) Graphs of the function $\varphi(t) := 1 + \varepsilon\left(t^2+\eta\right)^{-1/2}$ in Eq. (16) for several values of the parameter ε.

$t \in \mathbb{R}$, we get that the error between the l_1-norm and its approximation satisfies the bound $|\|\mathbf{a}\|_1 - \psi(\mathbf{a})| \leq (n\eta)^{1/2}$, $\mathbf{a} \in \mathbb{R}^n$. Indeed, the order of convergence of $\psi(\mathbf{a})$ to $\|\mathbf{a}\|_1$ is given by the number n of input points and the parameter $\eta^{1/2}$.

Replacing $\|\mathbf{a}\|_1$ with $\psi(\mathbf{a})$ in Eq. (2), we introduce the *smooth sparsification functional*

$$\mathcal{F}(\mathbf{a}) := \frac{1}{2}\left\| f - \sum_{i=1}^{n} a_i K(\mathbf{x},\mathbf{x}_i)\right\|_{\mathcal{H}}^2 + \varepsilon \sum_{i=1}^{n} (a_i^2+\eta)^{1/2} \tag{3}$$

and verify that its unique minimum solves a system of non-linear equations. To this end, we notice that the operator (3) is strictly convex because its summands $\|\cdot\|_{\mathcal{H}}$ and $s_\eta(t)$ are likewise strictly convex functions; in fact, the second order derivative $s''_\eta(t) = \eta(t^2+\eta)^{-3/2}$ is always positive. Then, the solution to the approximate sparsification problem $\arg\min_{\mathbf{a}\in\mathbb{R}^n}\{\mathcal{F}(\mathbf{a})\}$ is unique and satisfies the corresponding normal equation $\nabla\mathcal{F}(\mathbf{a}) = \mathbf{0}$.

Using the reproduction property $h(\mathbf{x}) = \langle K(\mathbf{x},\cdot), h\rangle_{\mathcal{H}}$, $h \in \mathcal{H}$, $\mathbf{x} \in \mathbb{R}^d$, the functional (3) is rewritten as

$$\mathcal{F}(\mathbf{a}) = \frac{1}{2}\|f\|_{\mathcal{H}}^2 - \sum_{i=1}^{n} y_i a_i + \frac{1}{2}\sum_{i,j=1}^{n} K(\mathbf{x}_i,\mathbf{x}_j) a_i a_j + \varepsilon\sum_{i=1}^{n}(a_i^2+\eta)^{1/2},$$

where $y_i := f(\mathbf{x}_i)$, $i = 1,\ldots,n$, is the set of f-values. Imposing $\nabla\mathcal{F}(\mathbf{a}) = \mathbf{0}$, the critical points of $\mathcal{F}$ satisfy the non-linear equations

$$\sum_{i=1}^{n} K(\mathbf{x}_i,\mathbf{x}_j) a_i + \varepsilon\frac{a_j}{(a_j^2+\eta)^{1/2}} = y_j, \quad j = 1,\ldots,n.$$

Indeed, the *sparsification normal equation* is

$$[K+\varepsilon\Delta(\mathbf{a})]\,\mathbf{a}=\mathbf{y},\quad \mathbf{a}:=(a_i)_{i=1}^n, \tag{4}$$

where $K:=(k_{ij})_{i,j=1}^n$, $k_{ij}:=K(\mathbf{x}_i,\mathbf{x}_j)$, is the Gram matrix associated to the input kernel, the diagonal matrix $\Delta(\mathbf{a}):=\operatorname{diag}((a_i^2+\eta)^{-1/2})_{i=1}^n$, involves the unknown coefficients, and $\mathbf{y}:=(y_i)_{i=1}^n$ is the right-hand side term of the f-values. Note that the symbol K indicates both the kernel and the corresponding Gram matrix. Finally, the entries of the Hessian matrix $H(\mathcal{F})$ of $\mathcal{F}$ are

$$H_{ij}(\mathcal{F})=\begin{cases} K(\mathbf{x}_i,\mathbf{x}_j) & i\neq j,\\ K(\mathbf{x}_j,\mathbf{x}_j)+\frac{\varepsilon\eta}{(a_j^2+\eta)^{3/2}} & \text{else,}\end{cases}$$

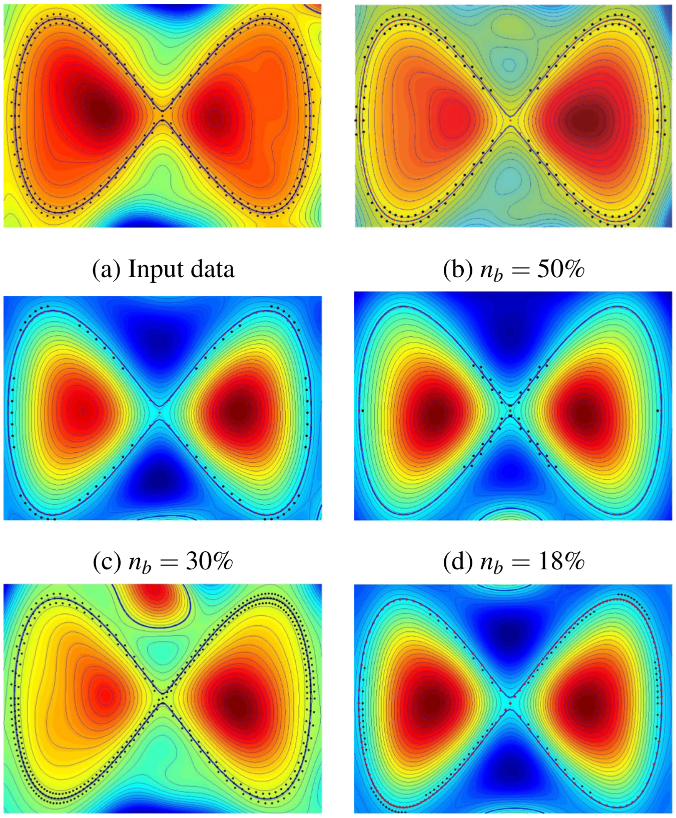
or, in matrix form,

$$H(\mathcal{F})=K+\varepsilon\eta\Delta^3(\mathbf{a}). \tag{5}$$

Since the matrices K and $\Delta^3(\mathbf{a})$ are positive-definite, the Hessian matrix (5) is positive definite for all $\mathbf{a}\in\mathbb{R}^n$, $\eta>0$, and $\varepsilon>0$. Then, the unique minimum of the strictly convex functional $\mathcal{F}$ is the solution to Eq. (4). We notice that the sparsification $K_\varepsilon:=K+\varepsilon\Delta(\mathbf{a})$ and Hessian matrices are evaluated in $O(n)$ time by updating only the diagonal entries of K and preserving its symmetry and sparsity. In particular, we bypass the approximation of differential operators with divided differences and related discretizations.

Sparse approximation of noisy data. To analyze the robustness of the sparsification scheme in case of noisy data, let us perturb the input data as $f(\mathbf{x}_i)=\tilde{y}_i$, $\tilde{y}_i:=y_i+e_i$, $i=1,\dots,n$, where $\tilde{y}_i$ is the measured value of f at $\mathbf{x}_i$ and e_i is a random variable with unknown probability distribution. Indicating with $\mathbf{e}:=(e_i)_{i=1}^n$ the error vector, the solutions of the noiseless and noisy problems satisfy the equations $[K+\varepsilon\Delta(\mathbf{a})]\,\mathbf{a}=\mathbf{y}$ and $[K+\varepsilon\Delta(\mathbf{b})]\,\mathbf{b}=\mathbf{y}+\mathbf{e}$. Using the upper bound

$$\|\Delta(\mathbf{a})\mathbf{a}\|_2=\left[\sum_{i=1}^n\frac{a_i^2}{a_i^2+\eta}\right]^{1/2}\leq n^{1/2}, \tag{6}$$



(a) Input data (b) $n_b = 50\%$

(c) $n_b = 30\%$ (d) $n_b = 18\%$

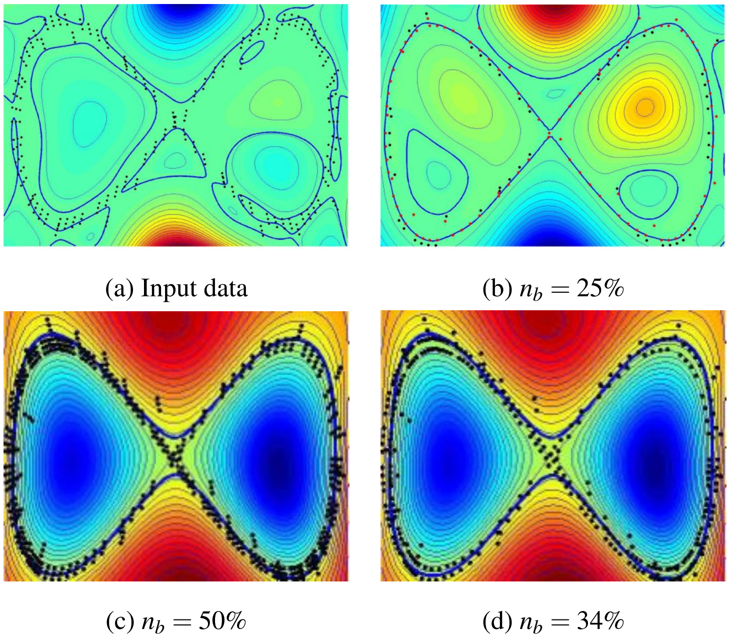
(e) Input data (f) $n_b = 23\%$

Figure 2: (a,e) Input data set (black dots) uniformly- and irregularly-sampled on and around the Bernoulli lemniscate $\mathcal{P} \subseteq \mathbb{R}^2$. (b-d, f) Reconstructed curves (blue line), selected centers, and iso-contours of the sparse approximations associated to (a,e), respectively. See also Fig. 3.

and the invertibility of the Gram matrix K, we get that the following relation holds

$$\begin{aligned}
\|\mathbf{a}-\mathbf{b}\|_2 &= \|K^{-1}\left[\varepsilon\left(\Delta(\mathbf{a})\mathbf{a}-\Delta(\mathbf{b})\mathbf{b}\right)+\mathbf{e}\right]\|_2\\
&\leq \|K^{-1}\|_2\left[\varepsilon\|\Delta(\mathbf{a})\mathbf{a}-\Delta(\mathbf{b})\mathbf{b}\|_2+\|\mathbf{e}\|_2\right]\\
&\leq \lambda_1^{-1}(K)\left[\varepsilon\left(\|\Delta(\mathbf{a})\mathbf{a}\|_2+\|\Delta(\mathbf{b})\mathbf{b}\|_2\right)+\|\mathbf{e}\|_2\right]\\
&\leq_{(6)} \lambda_1^{-1}(K)\left(2n^{1/2}\varepsilon+\|\mathbf{e}\|_2\right).
\end{aligned}$$

Therefore, the error $\|\mathbf{a}-\mathbf{b}\|_2$ is bounded by the inverse of the minimum eigenvalue of the Gram matrix K, the sparsification parameter ε, and the error magnitude $\|\mathbf{e}\|_2$. Preconditioning (if necessary) K, the resulting matrix has a generally low conditioning



(a) Input data

(b) $n_b = 25\%$

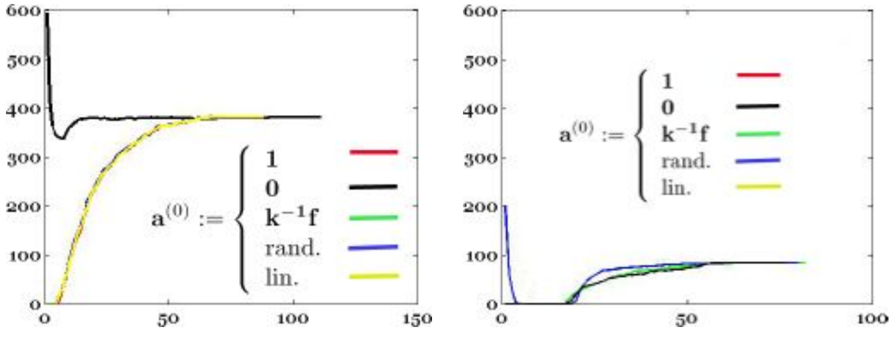
(c) $n_b = 50\%$

(d) $n_b = 34\%$

Figure 3: Sparse reconstructed curves using (a) all the input data (black dots) and (b) a set of selected centers. (c,d) Analogous examples on two different samplings (black dots) using the least-squares sparsification (Sect. 3.1).

number and the previous error bound is mainly guided by the error magnitude.

In Figs. 2, 3, we reconstruct the curve underlying a regularly- and irregularly-sampled noisy point set $\mathcal{P} := \{\mathbf{x}_i\}_{i=1}^{n}$ in $\mathbb{R}^2$ by computing the smooth and sparse function $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ that approximates the discrete map $f : \mathcal{P} \rightarrow \mathbb{R}$ such that $\{f(\mathbf{x}_i) = 0\}_{i=1}^{n}$. In order to avoid the trivial solution $f \equiv 0$, we add positive- and negative-valued normal constraints close to the null boundary conditions at $\mathbf{x}_i$ and in the directions $\mathbf{n}_i$ and $-\mathbf{n}_i$, where $\mathbf{n}_i$ is the normal at $\mathbf{x}_i$. These normal vectors are reliably estimated by applying the principal component analysis to the input data set $\mathcal{P}$ [44]. The shape of the sparse approximations confirms the stability of the sparsification scheme with respect to irregular samplings. Furthermore, the proposed approach provides both a sparse and smooth approximation of the input noisy data (Fig. 3) as a matter of the underlying Tikhonov regularization.



(a) (b)

Figure 4: With reference to Figs. 3 and 5, the graphs (a,b) show the sparsification induced by the iterative scheme with different initial guesses; on the x- and y-axis, we report the number of iterations and of null coefficients, respectively. The trade-off parameter is chosen as $\varepsilon = 0.5$.

2.2. Solving the sparsification normal equation

In the following, we discuss the iterative computation of the solution of the non-linear system (4), the choice of the initial guess, and the stop criteria. First of all, we derive a bound to the variation of the eigenvalues of the sparsification matrix $K_\varepsilon := K + \varepsilon\Delta(\mathbf{a})$, $\varepsilon > 0$, which will be used throughout the following discussion. Assuming that the eigenvalues of K are increasingly reordered and applying the Wielandt-Hoffman theorem concerning the eigenvalue sensitivity of symmetric matrices [23] (Ch. 8), we get that each eigenvalue $\lambda_i(K_\varepsilon)$ is related to the corresponding eigenvalues of K and $\Delta(\mathbf{a})$ by the bound

$$\lambda_i(K) + \varepsilon\lambda_1(\Delta(\mathbf{a})) \leq \lambda_i(K_\varepsilon) \leq \lambda_i(K) + \varepsilon\lambda_n(\Delta(\mathbf{a})). \tag{7}$$

Iterative solver. The solution to Eq. (4) is calculated by applying the *iterative sparsification scheme*

$$\left[K + \varepsilon\Delta(\mathbf{a}^{(r)})\right]\mathbf{a}^{(r+1)} = \mathbf{y} \leftrightarrow \mathbf{a}^{(r+1)} = \left[K + \varepsilon\Delta(\mathbf{a}^{(r)})\right]^{-1}\mathbf{y}, \tag{8}$$

with $\mathbf{a}^{(0)}$ initial guess. Let us now verify that the sequence $(\mathbf{a}^{(r)})_{r=0}^{+\infty}$ defined in Eq. (8) is bounded, thus guaranteeing its convergence to the solution of Eq. (4); in fact,

$$\begin{aligned}
\|\mathbf{a}^{(r)}\|_2 &= \left\| \left[K + \varepsilon\Delta(\mathbf{a}^{(r)})\right]^{-1} \mathbf{y} \right\|_2 \\
&\leq \lambda_1^{-1}\left(K + \varepsilon\Delta(\mathbf{a}^{(r)})\right) \|\mathbf{y}\|_2 \\
&\leq_{(7)} \left[\lambda_1(K) + \varepsilon\lambda_1(\Delta(\mathbf{a}^{(r)}))\right]^{-1} \|\mathbf{y}\|_2 \\
&\leq \lambda_1^{-1}(K)\|\mathbf{y}\|_2.
\end{aligned} \tag{9}$$

Since $\lambda_1(K) > 0$, the last term of the previous inequality is well-defined and independent of r. Therefore, a well-conditioned Gram matrix K, which is obtained through a pre-conditioner [23] (if necessary), guarantees the well-conditioning of K_ε. Furthermore, from the following upper bound to the conditioning number $\kappa_2(K_\varepsilon)$

$$\begin{aligned}
\kappa_2(K_\varepsilon) &= \frac{\lambda_n(K_\varepsilon)}{\lambda_1(K_\varepsilon)} \\
&\leq_{(7)} \frac{\lambda_n(K) + \varepsilon \max_{i=1,\ldots,n}\left\{(a_i^2+\eta)^{-1/2}\right\}}{\lambda_1(K) + \varepsilon \min_{i=1,\ldots,n}\left\{(a_i^2+\eta)^{-1/2}\right\}} \\
&\leq \frac{\lambda_n(K) + \varepsilon\eta^{-1/2}}{\lambda_1(K)} \\
&= \kappa_2(K) + \frac{\varepsilon\eta^{-1/2}}{\lambda_1(K)},
\end{aligned} \tag{10}$$

we get that the numerical stability of the iterative scheme in Eq. (8) is mainly controlled by the conditioning number $\kappa_2(K)$ and the eigenvalue $\lambda_1(K)$ of smallest magnitude. The conditioning number and the minimum eigenvalue of the matrix K are efficiently computed using iterative methods for the evaluation of the matrix spectrum [23] (Ch. 10). If K is ill-conditioned, then its preconditioning improves the conditioning of the matrix K_ε (c.f., Eq. (10)) and the computation stability. Our tests have shown that the matrices K and K_ε are generally well-conditioned.

Initial guess. The initial point $\mathbf{a}^{(0)}$ can be arbitrarily chosen or set equal to the *optimal guess* $\mathbf{a}_{opt}^{(0)} := K^{-1}\mathbf{y}$, which is the solution to the normal equation (4) with $\varepsilon := 0$. In this case, the difference between the solution $\mathbf{a}$ to Eq. (4) and $\mathbf{a}_{opt}^{(0)}$ is estimated as

$$\begin{aligned}\|\mathbf{a}-\mathbf{a}_{opt}^{(0)}\|_2 &= \varepsilon\|K^{-1}\Delta(\mathbf{a})\mathbf{a}\|_2 \\ &\leq_{(6)} \varepsilon\|K^{-1}\|_2\|\Delta(\mathbf{a})\mathbf{a}\|_2 \\ &\leq \varepsilon n^{-1/2}\lambda_1^{-1}(K).\end{aligned}$$

Indeed, a well-conditioned Gram matrix guarantees that the vector $\mathbf{a}_{opt}^{(0)} := K^{-1}\mathbf{y}$ provides a good initial guess of the iterative scheme. However, the computation of $\mathbf{a}_{opt}^{(0)}$ does not take into account the parameters ε, η and might be numerically unstable due to a possible ill-conditioning of the Gram matrix K.

To bypass these problems, we linearize Eq. (4) and use its solution as initial guess of the iterative scheme. Using the first-order Taylor polynomial $\psi(t) := \eta^{-1/2}t$ of the function $w_\eta(t) := t(t^2+\eta)^{-1/2}$, $t \to 0$, each entry $a_i(a_i^2+\eta)^{-1/2}$ of the vector $\Delta(\mathbf{a})\mathbf{a}$ is approximated by the linear term $\eta^{-1/2}a_i$, $i = 1,\dots,n$. Replacing $\Delta(\mathbf{a})\mathbf{a}$ with $\eta^{-1/2}\mathbf{a}$, Eq. (4) is approximated by the *linearized sparsification equation*

$$(K+\varepsilon\eta^{-1/2}I)\mathbf{a} = \mathbf{y}, \tag{11}$$

whose solution is used as initial guess $\mathbf{a}^{(0)}$ in Eq. (8). Note that the solution to Eq. (11) is efficiently computed through direct or iterative solvers of linear systems and is numerically stable, as a matter of the shift of its eigenvalues from $\lambda_i(K)$ to $\lambda_i(K)+\varepsilon\eta^{-1/2}$, $i = 1,\dots,n$.

To further analyze the dependence of the sparsified solution from the spectrum $\mathcal{S} := \{(\lambda_i(K),\mathbf{v}_i)\}_{i=1}^n$, $K\mathbf{v}_i = \lambda_i(K)\mathbf{v}_i$, of the Gram matrix K, let us rewrite the solution to the sparsification equation in terms of $\mathcal{S}$. Introducing the orthogonal eigenvector matrix $V := [\mathbf{v}_1,\dots,\mathbf{v}_n]$, $V^TV = VV^T = I$, and the diagonal eigenvalue matrix $\Delta := \mathrm{diag}(\lambda_i(K))_{i=1}^n$, we have $K = V\Delta V^T$ and the spectral representation of the solution to Eq. (11) is

$$\mathbf{a} = V(\Delta+\varepsilon\eta^{-1/2}I)^{-1}V^T\mathbf{y} = \sum_{i=1}^n \frac{1}{\lambda_i(K)+\varepsilon\eta^{-1/2}}\left(\mathbf{v}_i^T\mathbf{y}\right)\mathbf{v}_i. \tag{12}$$

Since the vector $\mathbf{a}$ is a linear combination of the eigenvectors of the Gram matrix and the filters are $\mu_i := (\lambda_i(K)+\varepsilon\eta^{-1/2})^{-1}$, $i = 1,\dots,n$, the behavior of $\mathbf{a}$ and its null entries are mainly controlled by the eigenvectors related to the lower eigenvalues. The effects of the choice of the initial guesses on the convergence and the number of iterations of the sparsification are discussed at the end of Section 2.3.

Sparseness and stop criteria. At the iteration r, we consider the entries of the coefficient vector $\mathbf{a}^{(r)}$ as null if their absolute values are lower than a given threshold σ. Our experiments have shown that $\sigma := 10^{-10}$ provides a good balance between numerical accuracy and sparsification percentage. Finally, the iteration stops when the number of null elements in $\mathbf{a}^{(r)}$ becomes stationary or when the residual error between two consecutive iterations is below a given threshold δ, i.e. $\|\mathbf{a}^{(r+1)} - \mathbf{a}^{(r)}\|_\infty \leq \delta$.

Approximation accuracy. To evaluate the approximation accuracy of the sparsification scheme, the values of the sparse approximation $g(\mathbf{x}) = \sum_{i=1}^{n} a_i K(\mathbf{x}, \mathbf{x}_i)$ on $\mathcal{P}$ are written in matrix form as $\mathbf{g} := (g(\mathbf{x}_i))_{i=1}^{n} = K\mathbf{a}$, where K is the Gram matrix and $\mathbf{a} := (a_i)_{i=1}^{n}$ is the solution to Eq. (4). Indicating with $\mathbf{y}$ the array of the f-values on $\mathcal{P}$, the least-squares approximation error between f and g on $\mathcal{P}$ is estimated as

$$\|\mathbf{y} - \mathbf{g}\|_2 = \varepsilon \|\Delta(\mathbf{a})\mathbf{a}\|_2 \leq_{(6)} \varepsilon n^{1/2}.$$

It follows that the approximation accuracy is proportional to the sparsification parameter ε and the value $\left[\sum_{i=1}^{n} \frac{a_i^2}{a_i^2 + \eta}\right]^{1/2}$, which depends on the number of null coefficients. This bound is also useful to tune the parameter ε with respect to the expected accuracy.

Computational cost. Assuming that the solution to the linear system (8) is computed with the conjugate gradient [23], the computational cost is $O(r(n + n\log n))$, where n and r, $r << n$, are the number of input data and steps of the iterative sparsification scheme, respectively. The solution of the sparsification system is also well-conditioned as a matter of the underlying regularization framework and based on a global sparsification procedure, which avoids time-consuming and *a-posteriori* local updates of the model. Finally, the input variables are not duplicated, thus reducing the memory allocation, which is one of the main drawbacks in case of a large amount of data.

2.3. Multi-resolutive sparse approximation

We now show that the number of coefficients of $\mathbf{a}^{(r)}$ that have been sparsified (i.e., considered as null) cannot decrease with respect to r, $r \geq 1$. Indeed, the iterative sparsification has an intrinsic *multi-resolutive structure*, which induces a sequence of nested approximation spaces. More precisely, let $\mathcal{H}_r$ be the linear space spanned by the functions $\{\varphi_i, i\in\mathcal{I}_r^C\}$, where $\mathcal{I}_r := \{i : a_i^{(r)} = 0\}$ is the set of corresponding null coefficients. For $r \geq s$, we will show that $\mathcal{I}_r \supseteq \mathcal{I}_s$ and therefore $(\mathcal{H}_r)_r$ is a sequence of nested spaces $\mathcal{H}_r \subseteq \mathcal{H}_s$ such that $g^{(r)} \in \mathcal{H}_r$.

Multi-resolutive structure. At each step $r \geq 1$, we consider the set $\mathcal{I}_r := \{i : a_i^{(r)} = 0\}$ of indices related to the functions that do not contribute to $g^{(r)}$; then, the sparse approximation $g^{(r)}$ at level r is $g^{(r)}(\mathbf{x}) = \sum_{i\in\mathcal{I}_r^C} a_i^{(r)}\varphi_i(\mathbf{x})$, where $\mathcal{I}_r^C$ is the complement of $\mathcal{I}_r$. From Eq. (8), it follows that

$$\sum_{j=1}^{n} k_{ij}a_j^{(r+1)} + \frac{\varepsilon}{\left(|a_i^{(r)}|^2+\eta\right)^{1/2}}a_i^{(r+1)} = y_i;$$

in particular, for $i \in \mathcal{I}_r$ we have

$$\sum_{j=1}^{n} k_{ij}a_j^{(r+1)} + \frac{\varepsilon}{\eta^{1/2}}a_i^{(r+1)} = y_i,$$

or equivalently,

$$a_i^{(r+1)} = \left(k_{ii} + \frac{\varepsilon}{\eta^{1/2}}\right)^{-1}\left(y_i - \sum_{j\neq i} k_{ij}a_j^{(r+1)}\right). \tag{13}$$

Being each component of $\mathbf{a}^{(r+1)}$ bounded (c.f., Eq. (9)), the second term of (13) is also bounded; assuming that $a_i^{(r)} \approx 0$ and using the relations $k_{ii} > 0$, $\eta << \varepsilon$, and $\eta \approx 0$, we obtain $a_i^{(r+1)} \approx 0$, $i \in \mathcal{I}_r$. Since $a_i^{(r)} \approx 0$ implies $a_i^{(r+1)} \approx 0$, $r \geq 1$, we conclude that the null entries of $\mathbf{a}^{(r)}$ are preserved in $\mathbf{a}^{(r+1)}$.

We further discuss the previous sparsification property from a numerical point of view; i.e., assuming that $|a_i^{(r)}| \leq \delta$, we estimate how much $|a_i^{(r+1)}|$ is close to zero. Rewriting the ith row of Eq. (4) as

$$\frac{\varepsilon}{(|a_i^{(r)}|^2+\eta)^{1/2}}a_i^{(r+1)} = y_i - \sum_{j=1}^{n} k_{ij}a_j^{(r+1)},$$

and normalizing the vector $\mathbf{y}$ to unitary l_2-norm, we get that

$$\begin{aligned}
\left|a_i^{(r+1)}\right| &= \left|\varepsilon^{-1}(|a_i^{(r)}|^2+\eta)^{1/2}\left[y_i-\sum_{j=1}^{n}k_{ij}a_j^{(r+1)}\right]\right| \\
&\leq \varepsilon^{-1}(|a_i^{(r)}|^2+\eta)^{1/2}\left|y_i-\sum_{j=1}^{n}k_{ij}a_j^{(r+1)}\right| \\
&\leq \varepsilon^{-1}(\delta^2+\eta)^{1/2}\|\mathbf{y}-K\mathbf{a}^{(r+1)}\|_2 \\
&\leq \varepsilon^{-1}(\delta^2+\eta)^{1/2}(\|\mathbf{y}\|_2+\|K\|_2\|\mathbf{a}^{(r+1)}\|_2) \\
&\leq_{(9)} \varepsilon^{-1}(\delta^2+\eta)^{1/2}\left[1+\frac{\lambda_n(K)}{\lambda_1(K)}\right] \\
&\leq \varepsilon^{-1}(\delta^2+\eta)^{1/2}\left[1+\kappa_2(K)\right].
\end{aligned}$$

Neglecting the constant term $\varepsilon^{-1}[1+\kappa_2(K)]$, which is assumed to be small by preconditioning K (if necessary), and recalling that $\eta \to 0^+$, the term $|a_i^{(r+1)}|$ is close to zero as much as the corresponding coefficient $|a_i^{(r)}|$ at the previous iteration, $r \geq 1$. This means that if $a_i^{(r)}$ is considered as null with respect to δ then $a_i^{(k)}$, $k \geq r$, is also treated as null. Indeed, we expect that the sparsity of each vector of the sequence generated by the iterative scheme increases until it converges. This property holds only for the sparsified entries of the vector $\mathbf{a}^{(r)}$ at a given iteration r; in fact, the initial vector $\mathbf{a}^{(0)}$ might be null or have some null entries (Fig. 4). For instance, this situation might happen with the initial condition $\mathbf{a}^{(0)} := K^{-1}\mathbf{f}$ and the linearized term (12). In these cases, we expect that the number of null entries decreases during the initial iterations and starts to increase when the sparsification starts to recognize redundant functions in the dictionary. Then, the number of null entries will grow until it becomes constant and the iteration stops. Our experiments have shown that this behavior is generally associated to the null initial condition $\mathbf{a}^{(0)} := \mathbf{0}$. With the initial guess $\mathbf{a}^{(0)} := \mathbf{1}$, the iterative sparsification scheme generally provides a strictly increasing number of null coefficients. Examples of sparsification curves with respect to different initial guesses are shown in Fig. 4; here, the choice of $\mathbf{a}^{(0)}$ slightly influences the number of iterations.

Convergence speed. To analyze the speed of convergence, we estimate the discrepancy between two consecutive steps of the iterative sparsification scheme (8). From the upper bound

$$\|\mathbf{a}^{(r+1)}-\mathbf{a}^{(r)}\|_2 = \varepsilon \left\| K^{-1}\left[\Delta(\mathbf{a}^{(r)})\mathbf{a}^{(r+1)} - \Delta(\mathbf{a}^{(r-1)})\mathbf{a}^{(r)}\right]\right\|_2 \leq 2\varepsilon n^{1/2}\lambda_1^{-1}(K),$$

it follows that the speed of convergence is generally higher in case of well-conditioned Gram matrices. Even though we cannot estimate the number of iterations, our experiments (e.g., Fig. 4) have shown that the number of iterations is generally small (e.g., lower than 50) and reduces by increasing the sparsification parameter ε.

3. Sparse approximations in arbitrary and discrete spaces

In the following, the iterative sparsification scheme is generalized to an arbitrary function space using radial basis functions and least-squares approximations, which replace the native norm in $\mathcal{H}$ with the l_2-norm (Sect. 3.1). Then, we introduce the discrete counterpart of the proposed sparsification scheme (Sect. 3.2) and discuss the main criteria for the selection of the parameters (Sect. 3.3).

3.1. Iterative least-squares sparse approximation in arbitrary spaces

The functional (3) involves a set of basis functions centered at each point of $\mathcal{P}$ and the approximation error has been measured with respect to the RKHS norm. If we deal with highly redundant dictionaries, or aim at further reducing the computational cost of the sparsification scheme, or select a set $\mathcal{B} := \{\varphi_i(\mathbf{x})\}_{i=1}^m$ of functions that are not generated by a kernel, then the error $\|f-g\|_{\mathcal{H}}$ in Eq. (2) is replaced by the least-squares constraint $\sum_{i=1}^n |y_i - g(\mathbf{x}_i)|^2$. Therefore, we introduce the *least-squares sparsification problem*

$$\arg\min_{g\in\mathcal{H}} \left\{ \frac{1}{2}\sum_{i=1}^n |y_i - g(\mathbf{x}_i)|^2 + \varepsilon\|\mathbf{a}\|_1 \right\}, \tag{14}$$

with $g(\mathbf{x}) := \sum_{i=1}^m a_i\varphi_i(\mathbf{x})$, $\mathbf{a} := (a_i)_{i=1}^m$, $m<n$. Here, each map $\varphi_i(\mathbf{x}) := \varphi(\|\mathbf{x}-\mathbf{c}_i\|_2)$, $i=1,\ldots,m$, is radially symmetric, generated by a map $\varphi : \mathbb{R}^+ \rightarrow \mathbb{R}$, and centered at a point of the set $\mathcal{C} := \{\mathbf{c}_i\}_{i=1}^m$, which is achieved by clustering the points of $\mathcal{P}$. The critical points of the smooth functional

$$\mathcal{F}(\mathbf{a}) := \frac{1}{2}\sum_{i=1}^n |y_i - g(\mathbf{x}_i)|^2 + \varepsilon\sum_{i=1}^n (|a_i|^2+\eta)^{1/2},$$

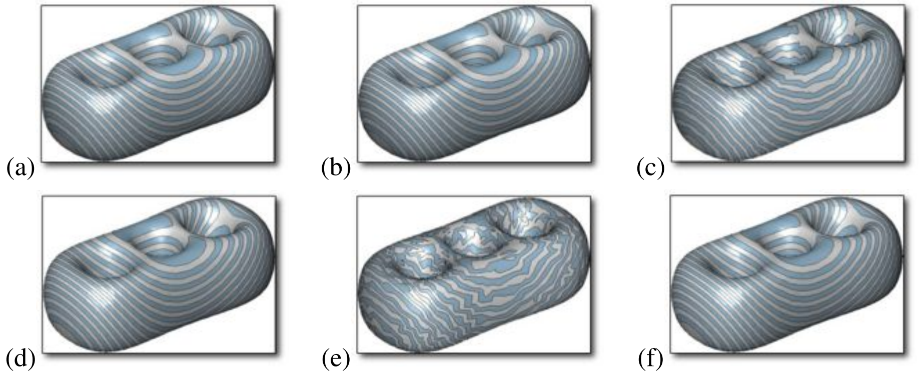


Figure 5: (a,c,e) Input and (b,d,f) least-squares sparse approximations of a scalar function $f:\mathcal{P}\rightarrow\mathbb{R}$ defined on a 3D shape $\mathcal{P}$ represented as a triangle mesh. Each sparse approximation uses the 27% of the input functions. The noise magnitude grows from (a) to (c) and (e). The dictionary includes 100 Laplacian eigenvectors and 500 randomly-generated maps. The corresponding sparsification function is shown in Fig. 4(b).

which approximates the objective function in Eq. (14), are the solutions to the system

$$\left[\tilde{K}^T\tilde{K}+\varepsilon\Delta(\mathbf{a})\right]\mathbf{a}=\tilde{K}^T\mathbf{y},\quad \tilde{K}:=(\varphi_j(\mathbf{x}_i))_{i=1,\dots,n}^{j=1,\dots,m},\quad \mathbf{y}:=(y_i)_{i=1}^{n},$$

whose dimension is $m\times m$ instead of $n\times n$. We notice that for each $\mathbf{a}\in\mathbb{R}^m$ and $\varepsilon>0$ the coefficient matrix $\left[\tilde{K}^T\tilde{K}+\varepsilon\Delta(\mathbf{a})\right]$ is positive-definite without assumptions on the matrix $\tilde{K}$. Furthermore, the discussion in Sect. 2.3 still applies to the least-squares sparsification scheme by substituting K with $\tilde{K}^T\tilde{K}$, $\mathbf{y}$ with $\tilde{K}^T\mathbf{y}$, and n with m. In this case, the vector $\mathbf{a}^{(0)}$ that solves the linearized sparsification problem

$$(\tilde{K}^T\tilde{K}+\varepsilon\eta^{-1/2}I)\mathbf{a}^{(0)}=\tilde{K}^T\mathbf{y}$$

converges to the least-squares solution $K^{\dagger}\mathbf{y}$ of the linear system $\tilde{K}\mathbf{a}^{(0)}=\mathbf{y}$, as $\eta\rightarrow 0^{+}$.

3.2. Discrete sparse approximations

In the Reproducing Kernel Hilbert Space $\mathcal{H}$, the scalar product is defined by the kernel function, and is related to a specific regularization operator [55, 56]. Assuming that the data is supported on a $(d-1)$-dimensional manifold, we derive a variant of the proposed approach that guarantees the smoothness of the solution in the ambient and

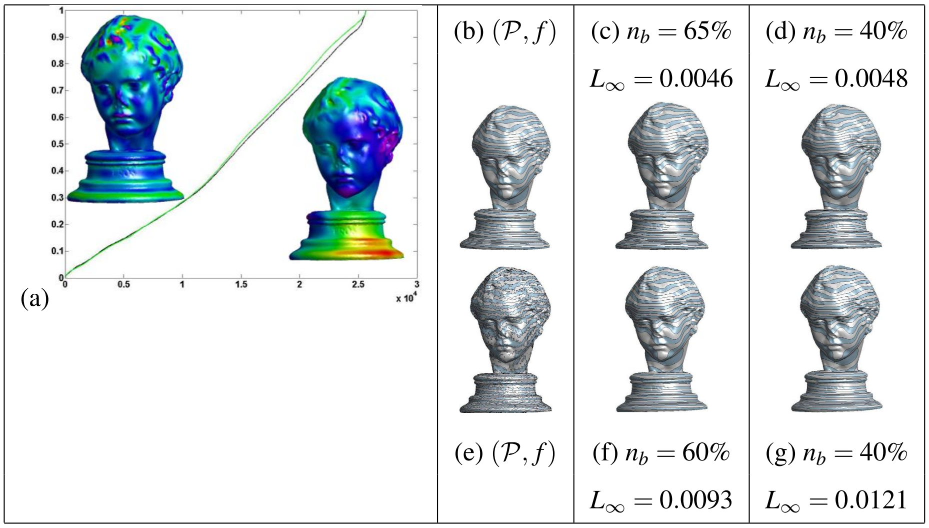


Figure 6: (a) Variation (y-axis) of the values assumed by the input (green line) and approximate (black line) map on $\mathcal{P}$ (x-axis). Level-sets of (b) a smooth, (e) noisy scalar function f defined on a 3D shape $\mathcal{P}$ and (c,d;f,g) corresponding sparse approximations (100 Laplacian eigenvectors and 300 random maps) achieved using a percentage n_b of maps selected in a dictionary of 400 maps defined on $\mathcal{P}$.

intrinsic space by using spectral graph theory and manifold learning techniques. Diagonalizing the Gram matrix of the sparsification normal equation, the unknown coefficients become independent; i.e., each non-linear equation involves only one unknown and its solution is computed in explicit form. In the discrete case, we introduce the sparsification scheme in a way similar to the continuous case. Since we deal with discrete data, any function $f : \mathcal{P} \rightarrow \mathbb{R}$ is uniquely identified by the vector $\mathbf{y} := (f(\mathbf{x}_i))_{i=1}^n$ of its values at the points of $\mathcal{P}$. Given the signal

$$\mathbf{g} = \sum_{i=1}^{m} a_i \mathbf{v}_i = V\mathbf{a} \in \mathbb{R}^n, \quad V := [\mathbf{v}_1, \dots, \mathbf{v}_m] \in \mathbb{R}^{n \times m}, \quad \mathbf{a} := (a_i)_{i=1}^m,$$

defined as a linear combination of the vectors in $\mathcal{B} := \{\mathbf{v}_i\}_{i=1}^m$, the functional (3) is replaced by

$$\mathcal{F}(\mathbf{a}) := \frac{1}{2}\|\mathbf{y} - V\mathbf{a}\|_S^2 + \varepsilon \sum_{i=1}^{n} (a_i^2 + \eta)^{1/2},$$

where the scalar product and the corresponding norm

$$\langle \mathbf{x}, \mathbf{y} \rangle_S := \mathbf{x}^T S \mathbf{y}, \qquad \|\mathbf{x}\|_S = \sqrt{\mathbf{x}^T S \mathbf{x}}, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

are induced by a $n \times n$ positive definite matrix S. The choice of S will be discussed later on. Since

$$\frac{1}{2}\|\mathbf{y}-V\mathbf{a}\|_S^2 = \frac{1}{2}\mathbf{y}^T\mathbf{y}-\mathbf{y}^T V\mathbf{a}+\frac{1}{2}\mathbf{a}^T V^T SV\mathbf{a},$$

the critical points of $\mathcal{F}$ satisfy the non-linear equations

$$\left[V^T SV + \varepsilon\Delta(\mathbf{a})\right]\mathbf{a} = V^T S\mathbf{y}. \tag{15}$$

Note that this expression is analogous to (4) with $K := V^T SV$ and $\mathbf{y} := V^T S\mathbf{y}$. Then, the solution to Eq. (15) is calculated by applying the iterative scheme

$$\left[V^T SV + \varepsilon\Delta(\mathbf{a}^{(r)})\right]\mathbf{a}^{(r+1)} = V^T S\mathbf{y} \Longleftrightarrow \mathbf{a}^{(r+1)} = \left[V^T SV + \varepsilon\Delta(\mathbf{a}^{(r)})\right]^{-1} V^T S\mathbf{y},$$

$r \geq 1$, with $\mathbf{a}^{(0)}$ initial guess.

Special case ($S := I$ and V orthogonal matrix). If $S := I$ is the identity matrix of order n, then Eq. (15) becomes $[I+\varepsilon\Delta(\mathbf{a})]\mathbf{a} = V^T\mathbf{y}$, whose coefficient matrix is diagonal. Therefore, each component a_i solves the equation

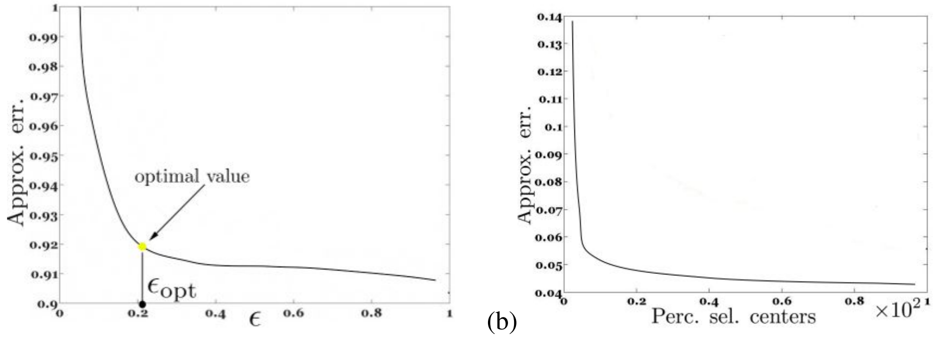
$$\left[1+\frac{\varepsilon}{(a_i^2+\eta)^{1/2}}\right]a_i = \mathbf{v}_i^T\mathbf{y}, \quad i = 1,\dots,m, \tag{16}$$

which involves only the unknown a_i.

General case. First of all, let us introduce locality relations among the points of the input data set $\mathcal{P} := \{\mathbf{x}_i\}_{i=1}^n$, through its *k-nearest neighbor graph* $\mathcal{T}$, where each point $\mathbf{x}_i \in \mathcal{P}$ is associated to the set $\{\mathbf{x}_j\}_{j\in\mathcal{N}_{\mathbf{x}_i}}$, $\mathcal{N}_{\mathbf{x}_i} \subseteq \{1,\dots,n\}$, of k-nearest points to $\mathbf{x}_i$, with respect to the Euclidean distance. Once $\mathcal{T}$ has been computed in $O(n\log n)$ time [4, 7], we consider the linear averaging operator $\mathbf{y} \mapsto L\mathbf{y}$, induced by the Laplacian matrix $L := (l_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}$,

$$l_{ij} := \begin{cases} -1 & i = j, \\ a_{ij}/\sum_{k\in\mathcal{N}_{\mathbf{x}_i}} a_{ik} & j \in \mathcal{N}_{\mathbf{x}_i}, \\ 0 & \text{else,} \end{cases}$$

with constant or Gaussian weights. For more details on the properties of the Laplacian matrix, we refer the reader to [5, 6, 12, 54]. Using the Laplacian matrix, we now specialize the proposed sparsification scheme to discrete data using spectral graph theory and manifold learning techniques.



(a) (b)

Figure 7: *L*-curves related to the examples in Figs. 5, 6. On the *y*-axis, we report the L_∞ approximation error between the input function and its sparsified approximation. On the *x*-axis, we report (a) the sparsification threshold ε and (b) the percentage of selected basis functions, respectively.

Choosing $S := S_1 + S_2$, where $S_1 := I$ and $S_2 := L$, the scalar product

$$\langle \mathbf{x}, \mathbf{y} \rangle_S := \mathbf{x}^T S \mathbf{y} = \mathbf{x}^T S_1 \mathbf{y} + \mathbf{x}^T S_2 \mathbf{y}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

is well-defined and the induced *discrete Sobolev norm* $\|\cdot\|_S$ measures the smoothness of the solution with respect to the *ambient space* and the *intrinsic geometry* [5, 13]. Alternatively, we select as S_1 the Gram matrix of a kernel function (e.g., the Gaussian kernel) and $S_2 := I$ or $S_2 := L$. The solution of the corresponding normal equation satisfies the relation

$$\left[V^T(S_1 + S_2)V + \varepsilon\Delta(\mathbf{a})\right]\mathbf{a} = V^T(S_1 + S_2)\mathbf{y}, \tag{17}$$

and is associated to the minimum $\mathbf{g} = \sum_{i=1}^m a_i \mathbf{v}_i$ of the functional

$$\mathcal{F}(\mathbf{a}) := \frac{1}{2}\left[\|\mathbf{y} - \mathbf{g}\|_{S_1}^2 + \|\mathbf{y} - \mathbf{g}\|_{S_2}^2\right] + \varepsilon \sum_{i=1}^m (a_i^2 + \eta)^{1/2}.$$

Computing the generalized eigendecomposition $S_2 V = S_1 V \Delta$ of the couple (S_1, S_2) [23], where V is the eigenvectors' matrix such that $V^T S_1 V = I$ and $\Delta := \mathrm{diag}(\lambda_i)_{i=1}^n$ is the eigenvalues' matrix, the normal equation (17) is rewritten as

$$[I + \Delta + \varepsilon\Delta(\mathbf{a})]\,\mathbf{a} = (I + \Delta)V^T S_1 \mathbf{y}.$$

Since the coefficient matrix is diagonal, each unknown a_i is the solution to the equation

$$\left[1 + \lambda_i + \frac{\varepsilon}{(a_i^2 + \eta)^{1/2}}\right] a_i = [1 + \lambda_i]\,\mathbf{x}_i^T S_1 \mathbf{y}, \quad i = 1, \ldots, n,$$

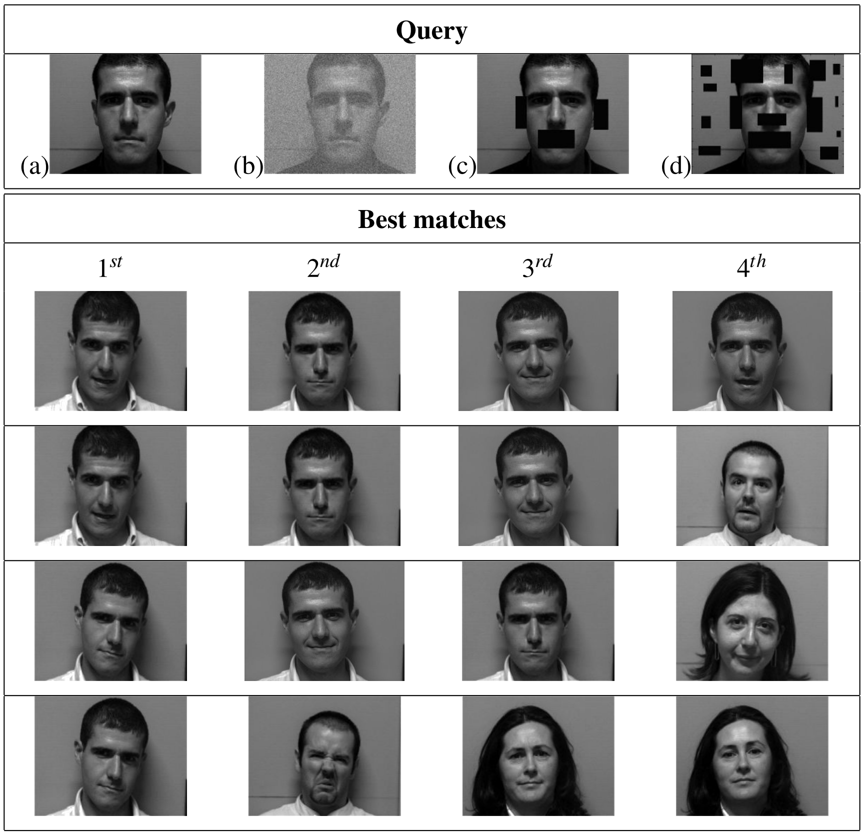


Figure 8: Face matching based on sparse approximations of the query images with respect to the dictionary provided by the input data set, which contains 144 face images. The input image in (a) has been degraded with (b) a Gaussian noise and (c,d) black boxes located on feature regions.

and it is computed through an iterative solver. Finally, we notice that the properties of the continuous sparsification scheme, discussed in Sect. 2, are also valid for the discrete approach by replacing the functions $\{\varphi_i(\mathbf{x})\}_{i=1}^{n}$ with the vectors $\{\mathbf{v}_i\}_{i=1}^{n}$ (e.g., the eigenvectors of L) and the signal f with the array of the f-values on $\mathcal{P}$.

In Fig. 5, we apply the discrete least-squares sparsification scheme to a discrete noisy signal defined on a 3-torus $\mathcal{P}$; here, the dictionary includes the Laplacian eigenvectors related to the first 100 eigenvalues of smaller magnitude and 500 discrete functions randomly generated using a Gaussian distribution on $\mathcal{P}$. The corresponding sparsification functions with different initial guesses of the iterative scheme are shown in Fig. 4(b). We notice that a different initial guess slightly affects the number of itera-

Method	Rate	Method	Rate	Method	Rate
PCA	84.65%	ICA	88.91%	NS	89.01%
L1	95.23%	L1-S	95.01%	=	=

Table 1: Verification rate at 0.1% false acceptance rate over a data set of 144 faces achieved by applying the Principal [59] (PCA) and Independent [29] Component Analysis (ICA); the Nearest Subspace [31] (NS); and the l^1-norm sparsification (L1) [61]; and the proposed smooth sparsification (L1-S).

tions. A similar example is presented in Fig. 6(b-g); in Fig. 6(a) the point-wise error related to the smooth (left) and noisy (right) scalar function is visualized using a color map that varies the hue component of the hue-saturation-value color model. The colors begin with red, pass through yellow, green, cyan, blue, magenta, and return to red. The l_∞-error between the input and sparsified approximation is lower than 10^{-4} (red).

3.3. Choice of the parameters

The positive constant ε in Eq. (3) controls the trade-off between the approximation error and the smooth sparsification term. As ε decreases (Fig. 7), the approximation error dominates the value of the functional $\mathcal{F}$; therefore, the solution is forced to precisely approximate all the f-values and the approximation error is minimized. As ε increases, the smoothness of the sparse approximation becomes predominant and filters out the local noise of f. Increasing ε, the iterative method converges to the null solution (i.e., *complete sparsification*) and generates the whole multi-resolutive scheme. To select the tradeoff ε between smoothness and approximation accuracy, statistical and heuristic methods (e.g., *L*-curve, best ration criterion) have been extensively discussed in [27, 60]. As general rule, we choose a value of ε enough big to achieve the whole sparsification and multi-resolutive approximation hierarchy $(\mathcal{H}_r)_r$ (Sect. 2.3); then, we consider the sparsification level r that provides a given sparsification rate or approximation accuracy. In our tests, the f-values have been normalized in $[0,1]$ and the full approximation hierarchy has been achieved with $\varepsilon := 1$. Finally, we have choosen $\eta := 10^{-14}$ and the initial guess of the iterative scheme has been set equal to the solution of the linearized sparsification equation (c.f., Eq. (11)).

According to [61], we apply the proposed smooth sparsification to face recognition.

To this end, the input data set $\mathcal{D}$ contains 144 face images with a 640×480 resolution, which represent 18 faces in 8 different poses (e.g., frontal, up, down, left view) and with different expressions (e.g., angry, disgusted). A representative image $\mathcal{I}_k$, $k=1,\ldots,18$, of each class of faces, which does not belong to $\mathcal{D}$, is degraded with Gaussian noise and with black boxes that cover several face features. Then, the resulting query images (4 faces for each class) are matched with the whole data set. For the comparison, the query image is represented as a sparse approximation of the the dictionary provided by $\mathcal{D}$ and the corresponding coefficients are used as image descriptors for matching. Fig. 8 shows the stability of the matching results with respect to the queries on a face image whose quality has been degraded in terms of smoothness and features. We notice that the image degradation introduces some wrong results in the query answers; however, the first matched images are always correctly recognized. The verification rate at 0.1% false acceptance rate is given in Table 1, which provides the comparison of the proposed smooth sparsification with the matching results provided by the Principal [59] and Independent [29] Component Analysis, the Nearest Subspace [31], and the l_1-norm sparsification [61]. We notice that the l_1-norm and smooth sparsification methods provide comparable results, which outperform previous work.

4. Conclusions and future work

This paper has discussed a multi-resolutive sparsification scheme, which is based on Tikhonov regularization and uses a smooth and strictly convex approximation of the l_1-norm. It differs from previous work for the use of an approximated formulation of the sparsification problem, which requires to solve a system of non-linear equations instead of applying heuristic solvers of convex quadratic optimization problems. Furthermore, it has a lower computational cost; is numerically stable as a matter of the underlying regularization framework; and provides a hierarchy of sparse approximations. We have also discussed its numerical properties, robustness to noise data, and specialization to a discrete space of functions using spectral graph theory and manifold regularization. As main application, we have considered the approximation of d-dimensional data. Even thought the number of selected functions is set in a simple way through a trade-off parameter, the main open issue is to control in an explicit way the sparsification percentage.

Acknowledgments

Special thanks are given to the anonymous Reviewers, whose comments helped us to improve the content of the paper. This work has been partially supported by the Research Project "*Methods and Techniques for the Development of Innovative Systems for Modeling and Analyzing Biomedical Data for Supporting Assisted Diagnosis*", PO CRO Programme, European Social Funding Scheme, Regione Liguria, and the FP7 Marie Curie Initial Training Networks *MultiScaleHuman* - "Multi-scale Biological Modalities for Physiological Human Articulation". The face data set is courtesy of Tristan Whitmar (Utrech Univerity, Netherlands) and the Aim@Shape Repository.

References

[1] Aharon, M., Elad, M., Bruckstein, A., 2006. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on Signal Processing 54 (11), 4311–4322.

[2] Andrew, G., Gao, J., 2007. Scalable training of l_1-regularized log-linear models. In: International Conference on Machine Learning. pp. 33–40.

[3] Aronszajn, N., 1950. Theory of reproducing kernels. Transactions of the American Mathematical Society 68, 337–404.

[4] Arya, S., Mount, D. M., Netanyahu, N. S., Silverman, R., Wu, A. Y., 1998. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. Journal of the ACM 45 (6), 891–923.

[5] Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15 (6), 1373–1396.

[6] Belkin, M., Niyogi, P., Sindhwani, V., 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research 7, 2399–2434.

[7] Bentley, J. L., 1975. Multidimensional binary search trees used for associative searching. Communication of the ACM 18 (9), 509–517.

[8] Candes, E. J., Demanet, L., Donoho, D., Ying, L., 2005. Fast discrete curvelet transforms. Multiscale Modeling and Simulation 5, 861–899.

[9] Candes, E. J., Emmanuel, J. C., Donoho, D. L., 1999. Curvelets-a surprisingly effective nonadaptive representation for objects with edges. Curves and Surfaces.

[10] Chen, S., Billings, A., Luo, W., 1989. Orthogonal least-squares methods and their applications to non-linear system identification. International Journal of Control 50, 1873–1896.

[11] Chen, S. S., Donoho, D. L., Saunders, M. A., 1998. Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computing 20 (1), 33–61.

[12] Chung, F. R. K., 1997. Spectral graph theory. American Mathematical Society.

[13] Coifman, R. R., Lafon, S., July 2006. Diffusion maps. Applied and Computational Harmonic Analysis 21 (1), 5–30.

[14] Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20 (3), 273–297.

[15] Donoho, D., Elad, M., Temlyakov, V., 2006. Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Transactions on Information Theory 52 (1), 6–18.

[16] Donoho, D., Huo, X., 2001. Uncertainty principles and ideal atomic decomposition. IEEE Transactions on Information Theory 47 (7), 2845–2862.

[17] Donoho, D. L., Elad, M., 2003. Optimally sparse representation in general (nonorthogonal) dictionaries via l-minimization. Proc. of the National Academy of Sciences 100 (5), 2197–2202.

[18] Efron, B., Hastie, T., Johnstone, L., Tibshirani, R., 2004. Least angle regression. Annals of Statistics 32, 407–499.

[19] Elad, M., Bruckstein, A. M., 2002. A generalized uncertainty principle and sparse representation in pairs of bases. IEEE Transactions on Information Theory 48, 2558–2567.

[20] Elhamifar, E., Vidal, R., 2009. Sparse subspace clustering. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2790–2797.

[21] Fuchs, J.-J., 2004. On sparse representations in arbitrary redundant bases. IEEE Transactions on Information Theory 50 (6), 1341–1344.

[22] Girosi, F., 1998. An equivalence between sparse approximation and support vector machines. Neural Computation 10 (6), 1455–1480.

[23] Golub, G., VanLoan, G., 1989. Matrix Computations. John Hopkins University Press, 2nd. edition.

[24] Goodman, J., 2004. Exponential priors for maximum entropy models.

[25] Gorodnitsky, I., Rao, B., 1997. Sparse signal reconstruction from limited data using focuss: a re-weighted minimum norm algorithm. IEEE Transactions on Signal Processing 45 (3), 600–616.

[26] Gribonval, R., M., N., 2003. Sparse decompositions in unions of bases. IEEE Transactions on Information Theory 49, 3320–3325.

[27] Hansen, P. C., O'Leary, D. P., 1993. The use of the L-curve in the regularization of discrete ill-posed problems. SIAM Journal of Scientific Computing 14 (6), 1487–1503.

[28] Holland, P. W., Welsch, R. E., 1977. Robust regression using iteratively reweighted least-squares. Communications in Statistics-theory and Methods 6, 813–827.

[29] Kim, J., Choi, J., Yi, J., Turk, M., 2005. Effective representation using ica for face recognition robust to local distortion and partial occlusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (12), 1977–1981.

[30] Le Pennec, E., Mallat, S., 2005. Sparse geometric image representations with bandelets. IEEE Transactions on Image Processing 14 (4), 423–438.

[31] Lee, K.-C., Ho, J., Kriegman, D., 2005. Acquiring linear subspaces for face recognition under variable lighting. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (5), 684–698.

[32] Lee, S.-I., Lee, H., Abbeel, P., Ng, A. Y., 2006. Efficient l_1 regularized logistic regression. In: Association for the Advancement of Artificial Intelligence.

[33] Lewicki, M. S., Olshausen, B. A., 1999. A probabilistic framework for the adaptation and comparison of image codes. Journal of the Optical Society of America A 16, 1587–1601.

[34] Lewicki, M. S., Sejnowski, T. J., 2000. Learning overcomplete representations. Neural Computation 12 (2), 337–365.

[35] Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A., 2008. Discriminative learned dictionaries for local image analysis. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8.

[36] Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A., 2008. Supervised dictionary learning. In: Neural Information Processing Systems. pp. 1033–1040.

[37] Mairal, J., Sapiro, G., Elad, M., 2008. Learning multiscale sparse representations for image and video restoration. Multiscale Modeling & Simulation 7 (1), 214–241.

[38] Mallat, S., Zhang, Z., 1992. Adaptive time-frequency decomposition with matching pursuits. Proc. of the Symposium on Time-Frequency and Time-Scale Analysis, 7–10.

[39] Mallat, S., Zhang, Z., 1993. Matching pursuit with time-frequency dictionaries. IEEE Transactions on Signal Processing 41, 3397–3415.

[40] Morse, B. S., Yoo, T. S., Chen, D. T., Rheingans, P., Subramanian, K. R., 2001. Interpolating implicit surfaces from scattered surface data using compactly supported radial basis functions. In: Proc. of Shape Modeling International and Applications. pp. 89–98.

[41] Ng, A. Y., 2004. Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In: ACM Proc. of the International Conference on Machine Learning. p. 78.

[42] Olshausen, B. A., Field, D. J., 1996. Natural image statistics and efficient coding. Network 7 (2), 333–339.

[43] Pati, Y., Rezaiifar, R., Krishnaprasad, P., 1993. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. Proc. of Signals, Systems and Computers, 40–44.

[44] Pauly, M., Gross, M., 2001. Spectral processing of point-sampled geometry. In: ACM Siggraph. pp. 379–386.

[45] Perkins, S., Theiler, J., 2003. Online feature selection using grafting. In: Proc. of the International Conference on Machine Learning. pp. 592–599.

[46] Peyré, G., Mallat, S., 2005. Surface compression with geometric bandelets. ACM Transactions on Graphics 24 (3), 601–608.

[47] Rao, B., Kreutz-Delgado, K., 1997. Deriving algorithms for computing sparse solutions to linear inverse problems. Proc. of Signals, Systems Computers 1, 955–959.

[48] Rao, B. D., Engan, K., Cotter, S. F., Palmer, J., Kreutz-delgado, K., 2003. Subset selection in noise based on diversity measure minimization. IEEE Transactions on Signal Processing 51 (3), 760–770.

[49] Rao, B. D., Kreutz-delgado, K., 1999. An affine scaling methodology for best basis selection. IEEE Transactions on Signal Processing 47 (1), 187–200.

[50] Rao, S., Tron, R., Vidal, R., Ma, Y., 2008. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8.

[51] Roth, V., 2004. The generalized lasso. IEEE Transactions on Neural Networks 15 (1), 16–28.

[52] Schoelkopf, B., Smola, A. J., 2002. Learning with Kernels. The MIT Press.

[53] Schoelkopf, B., Steinke, F., Blanz, V., 2005. Object correspondence as a machine learning problem. In: Proc. of the International Conference on Machine Learning. pp. 776–783.

[54] Sindhwani, V., Belkin, M., Niyogi, P., 2006. Geometric basis of semi-supervised learning. In: Semi-Supervised Learning. MIT Press, pp. 209–226.

[55] Smola, A. J., Schölkopf, B., Müller, K.-R., 1998. The connection between regularization operators and support vector kernels. Neural Networks 11 (4), 637–649.

[56] Steinke, F., Schölkopf, B., 2008. Kernels, regularization and differential equations. Pattern Recognition 41 (11), 3271–3286.

[57] Tipping, M. E., 2001. Sparse bayesian learning and the relevance vector machine. Journal of Machine Learning Reserach 1, 211–244.

[58] Tropp, J. A., 2006. Just relax: convex programming methods for identifying sparse signals in noise. IEEE Transactions on Information Theory 52 (3), 1030–1051.

[59] Turk, M., Pentland, A., Jun 1991. Face recognition using eigenfaces. IEEE Conference on Computer Vision and Pattern Recognition, 586–591.

[60] Wahba, G., 1990. Spline Models for Observational Data. Vol. 59. SIAM, Philadelphia.

[61] Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y., 2009. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2), 210–227.

[62] Yang, J., Wright, J., Huang, T., Ma, Y., 2008. Image super-resolution as sparse representation of raw image patches. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. pp. 1–8.