REDUCING THE DIMENSIONALITY OF DATA USING TEMPERED DISTRIBUTIONS

Rustem Takhanov
School of Sciences and Humanities
Nazarbayev University
53 Kabanbay Batyr Ave, Astana city
rustem.takhanov@nu.edu.kz

November 8, 2022

ABSTRACT

We reformulate unsupervised dimension reduction problem (UDR) in the language of tempered distributions, i.e. as a problem of approximating an empirical probability density function by another tempered distribution, supported in a $k$-dimensional subspace. We show that this task is connected with another classical problem of data science — the sufficient dimension reduction problem (SDR). In fact, an algorithm for the first problem induces an algorithm for the second and vice versa.

In order to reduce an optimization problem over distributions to an optimization problem over ordinary functions we introduce a nonnegative penalty function that “forces” the support of the model distribution to be $k$-dimensional. Then we present an algorithm for the minimization of the penalized objective, based on the infinite-dimensional low-rank optimization, which we call the alternating scheme. Also, we design an efficient approximate algorithm for a special case of the problem, where the distance between the empirical distribution and the model distribution is measured by Maximum Mean Discrepancy defined by a Mercer kernel of a certain type. We test our methods on four examples (three UDR and one SDR) using synthetic data and standard datasets.

Keywords: linear dimensionality reduction, sufficient dimension reduction, alternating scheme, tempered distribution.

1 Introduction

Linear dimension reduction (LDR) is a family of problems in data science that includes principal component analysis, factor analysis, linear multidimensional scaling, Fisher’s linear discriminant analysis, canonical correlations analysis, sufficient dimension reduction (SDR), maximum autocorrelation factors, slow feature analysis and more. In unsupervised dimension reduction (UDR) we are given a finite number of points in $\mathbb{R}^n$ (sampled according to some unknown distribution) and the goal is to find a “low-dimensional” manifold (e.g. an affine or a linear subspace) that approximates “the support” of the distribution. UDR, historically, was approached by linear methods and, therefore, has developed into a set of standard tools in data science. Though non-linear dimensionality reduction (a.k.a. the manifold learning) techniques gained a wide popularity in modern research, the potential of linear methods is far from being exhausted. For high-dimensional datasets, due to the phenomenon of concentration of measure \cite{1}, LDR often can give us an interpretable and low-dimensional representation of data. The linearity of a projection operator is a restrictive property that allows avoiding the overfitting in the dimension reduction (which is a key problem for the manifold learning techniques).

The LDR study field currently achieved a saturation level at which unifying frameworks for the problem become of special interest. One of such frameworks, that covers many cases of LDR, frames LDR as the optimization task over matrix manifolds such as the Stiefel manifold \cite{2}. Elements of the Stiefel manifold $V_k(\mathbb{R}^n)$ are orthogonal $k$-frames $O \in \mathbb{R}^{n \times k}$ whose column space is the $k$-dimensional space onto which a dataset is projected. Different loss functions on $V_k(\mathbb{R}^n)$ define different versions of LDR. Table 1 of \cite{2} lists fourteen common LDR techniques (such as principal component analysis, multi-dimensional scaling, linear discriminant analysis etc), nine of which...
are formulated over Stiefel manifolds. Such a general treatment allows to approach all LDR problems by a single
algorithm, i.e. by an adaptation of the gradient descent method to Stiefel manifolds \cite{3}. This adaptation consists
of a series of projected gradient steps where a common gradient descent is followed by a projection onto a Stiefel
manifold, which is equivalent to the computation of a singular value decomposition of a current point. Note that the
Stiefel manifold is a non-convex set and even minimizing a convex function on such a manifold is an NP-hard task, in
general. Although, in applications, the projected gradient descent method demonstrated a relatively fast convergence
to good quality solutions.

The paper’s main contribution is a development of a novel view of LDR. First, an argument over which we search
in an optimization task is not a \(k\)-frame, but a tempered distribution (which is a generalization of a probabilistic
distribution) that is concentrated on a \(k\)-dimensional linear subspace of \(\mathbb{R}^n\). Thus, an argument has a more complex
structure, it includes not just a \(k\)-dimensional subspace, but also a distribution on that subspace. The justification
of our optimization framework uses the theory of generalized functions, or tempered distributions \cite{4, 5}. An important
generalized function that cannot be represented as an ordinary function is the Dirac delta function, denoted \(\delta\), and \(\delta^n\)
denotes its \(n\)-dimensional version.

This more general formulation allows us to analyze new types of objectives for LDR. In Section 4 we list four examples
of such objectives that, to our knowledge, have not been considered in the LDR field so far. A notable specifics of
such objectives is that, even for a fixed \(k\)-dimensional subspace \(L\), finding an optimal distribution supported in \(L\) is
a non-trivial optimization task. In other words, our problems can not be simply reduced to the previous formalisms
based on the Stiefel manifold, or the Grassmannian \cite{6, 7}.

Let us briefly describe an optimization problem that motivates the new formalism. Any dataset \(\{x_i\}_{i=1}^N \subseteq \mathbb{R}^n\) naturally
corresponds to the distribution

\[
p_{\text{emp}}(x) = \frac{1}{N} \sum_{i=1}^N \delta^n(x - x_i)
\]

which, with some abuse of terminology, can be called the empirical probability density function. Based on that, UDR
can be understood as a task whose goal is to approximate \(p_{\text{emp}}(x)\) by \(q(x)\), where \(q(x)\) is a distribution whose density
is supported in a \(k\)-dimensional linear subspace \(L \subseteq \mathbb{R}^n\). Note that a function whose density is supported in some
low-dimensional subset of \(\mathbb{R}^n\) is not an ordinary function. An exact definition of a set of such distributions, denoted
by \(\mathcal{G}_k\), is given in Section 3. To formulate an optimization task we additionally need a loss \(D(p_{\text{emp}}, q)\) that measures
the distance between the ground truth \(p_{\text{emp}}\) and a distribution \(q\), that we search for. Thus, in our approach, the UDR
problem is defined as

\[
I(q) = D(p_{\text{emp}}, q) \rightarrow \min_{q \in \mathcal{G}_k}
\]

under the condition that \(q(x)\) has a \(k\)-dimensional support. In most of our statements we do not consider any specific
loss functions, though in our basic examples we deal with the Maximum Mean Discrepancy distance or the Wasserstein
distance.

**The UDR and SDR.** Within our formalism the sufficient dimension reduction problem is tightly connected with the
UDR problem. In the SDR, given supervised data, the goal is to find the so called effective subspace, defined
by its orthogonal basis (or, a \(k\)-frame) \(\{w_1, \cdots, w_k\} \subseteq \mathbb{R}^n\), such that the regression function can be searched in
the form \(g(w_1^T x, \cdots, w_k^T x)\). In literature, these functions are known under different names, e.g. functions with
low effective dimensionality \cite{8}, functions with active subspaces \cite{9} and multi-ridge functions \cite{10, 11}. In \cite{12}
it was shown that a method originally developed for the SDR can be turned into a UDR method, i.e. applied to
unsupervised data, by simply setting an output to be equal to an input. In such methods for the SDR problem as the
Sliced Inverse Regression \cite{13}, the Principal Hessian Direction \cite{14}, the Sliced Average Variance Estimation \cite{15}, an
effective subspace is recovered from the Singular Value Decomposition applied to a certain matrix that is constructed
from a training set in a straightforward way. Other methods, such as the Principal Fitted Components \cite{16}, the
Likelihood Acquired Direction \cite{17}, the Kernel Dimensionality Reduction \cite{18}, are based on analytic expressions
measuring the affinity of a \(k\)-dimensional subspace to the effective subspace. The second type of methods reduce the
SDR problem to an optimization problem over the Stiefel manifold, or the Grassmanian. For other methods we refer
to a tutorial on SDR methods \cite{19}. Again, an important aspect of all these methods is that, given a fixed effective
subspace, the regression function that predicts an output variable has a relatively straightforward structure and is not
optimized by any additional supervised learning procedure. The key novelty that our framework brings to the SDR is
that we suggest to search for an effective subspace and a regression function in a joint manner.

The key observation of our analysis, stated in Theorem 2, is that a class of functions of the form \(g(w_1^T x, \cdots, w_k^T x)\)
can be characterized as functions whose Fourier transform is supported in the corresponding effective subspace. In
other words, functions with an effective dimensionality \(k\) are dual to \(\mathcal{G}_k\) under the Fourier transform. Three examples
of UDR problems that we give in Section 4 are cast as (2), whereas in the fourth example we formulate SDR as
an optimization task with the search space dual to that of UDR (to distinguish our formulation from a general SDR problem we call it an SDR with optimized regression function). Thus, all four examples can be studied within our optimization framework.

Besides the problem setup we also suggest a general algorithm that tackles it. The basic idea of that algorithm, which we call the alternating scheme, instead of optimizing over $G_k$, to optimize over ordinary functions with a penalty added to an objective that forces the ordinary function’s support to be low-dimensional.

The penalty based reformulation. The starting point of our approach is to reduce the task (2) to the minimization of $I(q) + \lambda R(q)$ over ordinary functions $q$. We define the penalty function $R(q)$ in such a way that forcing $R(q)$ to be small is equivalent to forcing “the support” of $q$ to be $k$-dimensional. Our definition of $R$ is based on using a positive definite kernel $M : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{C}$.

First we note that $M$ defines a bilinear form on pairs of (possibly, generalized) functions by $(f|M|g) = \int_{\mathbb{R}^n \times \mathbb{R}^n} f(x)^* M(x,y) g(y) dx dy$. On a properly defined space of (generalized) functions, the bilinear form $(\cdot|M|\cdot)$ is the hermitian inner product, using which one can define distances and other geometrical notions on that space. Note that if $f$ and $g$ are probability density functions and $M$ is real-valued, the corresponding distance function, i.e. $\text{dist}_M(f,g) = ((f|M|f) - (g|M|g))/2$ coincides with the maximum mean discrepancy metric [20]. We define $R(q)$ as

$$R(q) = \sum_{i=k+1}^n \lambda_i(M(q))$$

where $M(q) = \text{Re}[\langle x_i q(x) | M | x_i q(x) \rangle]$, and $\lambda_1(M(q)) \geq \lambda_2(M(q)) \geq \cdots$ are ordered eigenvalues of the matrix $M(q)$. The sum of all but first $k$ eigenvalues of a positive semidefinite matrix $A$ is a well-known penalty function, denoted by $\|A\|_{n-k}$, and called a Ky Fan $n-k$-antinorm. Applications of the Ky Fan $n-k$-antinorm to low-rank optimization problems can be found in [21] [22] [23] and its properties are studied in [24].

Thus, we reduce the task (2) to

$$I(q) + \lambda \|M(q)\|_{n-k} \to \min_q$$

over ordinary functions. An analysis that we make in Subsection 5.3 of Section 3 (based on theory of tempered distributions) shows that if the kernel $M$ is chosen from a class of so called proper kernels and the solution of (2) satisfies certain regularity conditions, the solution of the task (4) for $\lambda \to +\infty$ will approach the solution of (2).

The alternating scheme. The task (4) can be understood as an infinite dimensional low-rank optimization task in which the penalty term forces the matrix $M(q)$ to be of rank $k$. In Section 6 we prove that $M(q) = S_q S_q^*$ where $S_q$ is a linear operator between a suitable space $H_q$ and $\mathbb{R}^n$ that itself depends on $q$ linearly, and this automatically gives us that $R(q) = \min_{S_q} \|S_q - S_q^*\|^2_*$ where the minimum is taken over all operators between $H_q$ and $\mathbb{R}^n$ of rank $k$ and $\|\cdot\|_*$ is a suitable norm on the space of bounded linear operators from $H_q$ to $\mathbb{R}^n$.

Then, a natural idea to solve the task (4) is to present it as the joint minimum $\min_q \min_{S: \text{rank} S = k} I(q) + \lambda \|S_q - S\|^2_*$ and to minimize the objective over $q$ and over $S$ of rank $k$ in an alternating fashion, i.e.

$$q_{l+1} = \text{arg} \min_q I(q) + \lambda \|S_q - S_l\|^2_*$$

$$S_{l+1} = \text{arg} \min_{S: \text{rank} S = k} \|S_{q_{l+1}} - S\|^2_*$$

This algorithm, called the alternating scheme, is suitable for a practical implementation due to the fact that the second step of it, i.e. the optimization over $S$ of rank $k$, is solvable analytically. In fact, $S_{l+1}$ is the Singular Value Decomposition of $S_{q_{l+1}}$ truncated at $k$-th term. In Section 8 we give an algorithm whose every step is equivalent to a corresponding step of the alternating algorithm, but it operates on Fourier transforms of functions rather than on functions of initial coordinates. Numerical specifications of the alternating scheme for different special cases of UDR/SDR problems are given in [8], [11] and [14]. In Section 8 we describe results of our experiments with the alternating scheme that we conducted for various synthetic and practical datasets. As a result we conclude that the alternating scheme is a practical algorithm that can be applied to datasets of moderate size. For the SDR tasks its performance is comparable with classical algorithms.

An approximate algorithm for a special case. For a special case of the task (2), where the distance function is the Maximum Mean Discrepancy with the kernel of the form $K(x,y) = (x \cdot y)H(x,y)$ and $H$ is itself a Mercer kernel, we develop an approximate algorithm that can be applied to large datasets. In Section 7 we demonstrate that a solution with provable approximation ratios is given by the following simple procedure: given a dataset $\{x_i\}_{i=1}^N$ we build a data matrix $X = [x_1, \cdots, x_N]$, a Gram matrix $G = [H(x_i, x_j)]$, and output first $k$ principal components of
the matrix $XGXT$. This algorithm is tested on Yale B dataset for the shadow/black removal and SBMnet datasets for the background modeling. In both applications, our approximate algorithm showed a performance comparable to the performance of other low-rank approximation algorithms.

The structure of the paper is as follows. In Section 2 we give some notations and define standard notions from functional analysis that we use throughout the paper. In Section 3, we formally define the search space in Problem 1, denoted $G_k$, and an image of $G_k$ under the Fourier transform, denoted $F_k$. In Section 4, we formulate some UDR/SDR problems as optimization tasks over $G_k/F_k$. Instead of searching directly in a set of generalized functions, $G_k$, in Section 5, we describe how we substitute an ordinary function for a distribution in the optimization task at the expense of adding a new penalty term to its objective, $\lambda R(f)$. Using a kernel $M(x,y)$, Theorem 4 characterizes generalized $g \in G_k$, as such $g$ for which the matrix of properly defined integrals $M_g = \text{Re} \left[ \int_{\mathbb{R}^n} \times_{ij} f(x)^* M(x,y) g(y)dx dy \right]_{i,j=1}^\infty$ is of rank $k$. In Section 6, we suggest a method for solving $\min_{\phi} I(\phi) + \lambda R(\phi)$ which we call the alternating scheme. In Section 7, we describe a simple approximate algorithm for the task (2) in a special subcase of the Maximum Mean Discrepancy distance and prove some theoretical guarantees on the approximation ratio of this algorithm. Section 8 is dedicated to experiments with the alternating scheme on synthetic and real world data and with the approximate algorithm on the shadow/black removal and the background modeling applications. Proofs of all theorems and lemmas are given after their formulations, or can be found in the appendix.

1.1 Related work

As was already mentioned, another unifying framework for LDR tasks is suggested by [2] in which the basic search space is the Stiefel manifold $V_k(\mathbb{R}^n)$. The main advantage of the Stiefel manifold over $G_k$ is that its elements are finite-dimensional. Because a distribution from $G_k$ is an infinite-dimensional object, an optimization over $G_k$ requires additional constructions to turn it into a finite-dimensional task. Both an optimization over $G_k$ and over $V_k(\mathbb{R}^n)$ is typically hard: for a final point, at best one can guarantee that it is a local extremum. Promising aspects of $G_k$ are: a) $G_k$ allows to formulate a new class of objectives naturally on it, b) local extrema on $G_k$ substantially differ from local extrema on $V_k(\mathbb{R}^n)$, because a local search over $G_k$ uses more degrees of freedom.

There is plenty of literature on the SDR problem some of which was already mentioned. In [20], the Fourier transform was applied for estimating the effective subspace in SDR, implicitly using an analog of Theorem 2. The closest to ours is a recent approach of [27], where an effective subspace was computed in a two step process. First, given supervised data, a regression function was trained in the form of a neural network (with a general architecture), then the obtained regression function was approximated by another neural network with a bottleneck architecture (by which a low effective dimensionality is guaranteed by construction). Like in this approach, we train a regression function as a neural network, though we search for it and an effective subspace jointly. In our approach, it is a regularization term $R(f)$ that forces the neural network to have a low effective dimensionality.

Using Ky Fan $k$-antinorm as a regularizer for the matrix completion problem has been suggested by [21] and further developed in [22, 23, 24]. Unlike this chain of works, we formulate an infinite-dimensional task and our regularizer $R(f) = \|M_f\|_{ln-k}$ is a sum of smallest $n - k$ squared singular values of the infinite-dimensional operator $S_f$ where $S_f$ depends on $f$ linearly and $M_f = S_f S_f^T$. Thus, our algorithms are substantially different from algorithms designed within the latter approach. The idea of alternating two basic stages, convex optimization and SVD, is ubiquitous in low-rank optimization, see e.g. [25, 29].

2 Preliminaries and notations

Throughout this paper we use standard terminology and notation from functional analysis. For details one can address the textbook on the theory of distributions [30]. The Schwartz space, denoted by $S(\mathbb{R}^n)$, is a space of infinitely differentiable functions $f : \mathbb{R}^n \to \mathbb{C}$ such that $\forall \alpha, \beta \in \mathbb{N}^n, \sup_{x \in \mathbb{R}^n} |x^\alpha D^\beta f(x)| < \infty$, and equipped with standard topology. Its dual space is denoted by $S'(\mathbb{R}^n)$ and is equipped with weak topology. For a tempered distribution $T \in S'(\mathbb{R}^n)$ and $\phi \in S(\mathbb{R}^n)$, $\langle T, \phi \rangle$ denotes $T(\phi)$. Thus, for a sequence $\{f_s\} \subseteq S'(\mathbb{R}^n)$ and $f \in S'(\mathbb{R}^n)$, $\lim_{s \to \infty} f_s = f$ (or, $f_s \to f^*$) means that $\lim_{s \to \infty} \langle f_s, \phi \rangle = \langle f, \phi \rangle$ for any $\phi \in S(\mathbb{R}^n)$. For a sequence $\{f_s\}_{s=1}^\infty \subseteq S'(\mathbb{R}^n)$, $\lim_{s \to \infty} f_s$ denotes a set of points $f \in S'(\mathbb{R}^n)$, such that there exists a growing sequence $\{s_i\} \subseteq \mathbb{N}$ and $\lim_{i \to \infty} f_{s_i} = f$. The Fourier and inverse Fourier transforms are denoted by $\mathcal{F}, \mathcal{F}^{-1} : S'(\mathbb{R}^n) \to S'(\mathbb{R}^n)$. For brevity, we denote $\mathcal{F}[f]$ by $\hat{f}$. If all required conditions are satisfied, an integrable $f : \mathbb{R}^n \to \mathbb{C}$ (or, a Borel measure $\mu$ on $\mathbb{R}^n$) is used as the tempered distribution $T_f$ (or, $T_{\mu}$) where $(T_f, \phi) = \int_{\mathbb{R}^n} f(x)\phi(x)dx$ (or, $(T_{\mu}, \phi) = \int_{\mathbb{R}^n} \phi(x)d\mu$). For $\Omega \subseteq S'(\mathbb{R}^n)$, $\overline{\Omega}$ denotes the sequential closure of $\Omega$ with respect to standard topology of $S'(\mathbb{R}^n)$. For $\Omega \subseteq S'(\mathbb{R}^n)$, $\overline{\Omega}$ denotes the sequential closure of $\Omega$ with respect to weak topology of $S'(\mathbb{R}^n)$. For $\psi \in S(\mathbb{R}^n), T \in S'(\mathbb{R}^n)$, the
convolution is defined as a tempered distribution \( \psi * T \) such that \( \langle \psi * T, \phi \rangle = \langle T, \widetilde{\psi} * \phi \rangle \) where \( \tilde{\psi}(x) = \psi(-x) \). If \( T \in \mathcal{S}'(\mathbb{R}^n) \) and a function \( \psi \) is such that \( \psi(x)\phi(x) \in \mathcal{S}(\mathbb{R}^n) \) whenever \( \phi(x) \in \mathcal{S}(\mathbb{R}^n) \), then the multiplication \( \psi T \) is defined by \( \langle \psi T, \phi \rangle = \langle T, \psi \phi \rangle \). Given a measure \( \mu \), by \( L_{2,\mu}(\mathbb{R}^n) \) we denote the complex \( L_2 \)-space with the inner product \( \langle u, v \rangle_{L_{2,\mu}} = \int u(x)^* v(x) d\mu \). The induced norm is then \( \|u\|_{L_{2,\mu}} = \sqrt{\langle u, u \rangle_{L_{2,\mu}}} \). If \( dx_0 = p(x) dx \), then \( L_{2,\mu} \) is denoted by \( L_{2,p} \). A set of infinitely differentiable functions in \( \mathbb{R}^n \) is denoted by \( C^\infty(\mathbb{R}^n) \). A set of infinitely differentiable functions with compact support in \( \mathbb{R}^n \) is denoted by \( C_c^\infty(\mathbb{R}^n) \). If \( T \) is a topological space, then a subset \( S \subseteq T \) is said to be dense in \( T \) if the sequential closure of \( S \) is equal to \( T \). For a square matrix \( A \), \( \text{Tr}(A) \) denotes its trace and for an arbitrary matrix, \( A \|_F \overset{\text{def}}{=} \sqrt{\text{Tr}(A^T A)} \). The identity matrix of size \( n \) is denoted by \( I_n \). The notation \( f \prec g \) means \( f = cg \) where \( c \) is some universal constant.

### 3 Basic function classes

To formalize distributions supported in a \( k \)-dimensional subspace, we need a number of standard definitions [31]. For \( \phi_1 \in \mathcal{S}(\mathbb{R}^k) \) and \( \phi_2 \in \mathcal{S}(\mathbb{R}^{n-k}) \), their tensor product is the function \( \phi_1 \otimes \phi_2 \in \mathcal{S}(\mathbb{R}^n) \) such that \( (\phi_1 \otimes \phi_2)(x, y) = \phi_1(x)\phi_2(y) \). The span of \( \{ \phi_1 \otimes \phi_2 \mid \phi_1 \in \mathcal{S}(\mathbb{R}^k), \phi_2 \in \mathcal{S}(\mathbb{R}^{n-k}) \} \), denoted by \( \mathcal{S}(\mathbb{R}^k) \otimes \mathcal{S}(\mathbb{R}^{n-k}) \), is called the tensor product of \( \mathcal{S}(\mathbb{R}^k) \) and \( \mathcal{S}(\mathbb{R}^{n-k}) \). For \( g_1 \in \mathcal{S}(\mathbb{R}^k) \) and \( g_2 \in \mathcal{S}(\mathbb{R}^{n-k}) \), their tensor product is defined by the following rule: \( (g_1 \otimes g_2)(\phi_1, \phi_2) = (g_1, \phi_1)(g_2, \phi_2) \) for any \( \phi_1 \in \mathcal{S}(\mathbb{R}^k), \phi_2 \in \mathcal{S}(\mathbb{R}^{n-k}) \). Since \( \mathcal{S}(\mathbb{R}^k) \otimes \mathcal{S}(\mathbb{R}^{n-k}) = \mathcal{S}(\mathbb{R}^n) \), there is only one distribution \( g_1 \otimes g_2 \in \mathcal{S}(\mathbb{R}^n) \) that satisfies the identity.

An example of a generalized function, whose density is concentrated in a \( k \)-dimensional subspace, is any distribution that can be represented as \( g \otimes \delta^{n-k} \overset{\text{def}}{=} g \otimes \delta \otimes \cdots \otimes \delta \) where \( g \in \mathcal{S}(\mathbb{R}^k) \). If \( g = T_f \), where \( f : \mathbb{R}^k \to \mathbb{R} \) is an ordinary function, then \( g \otimes \delta^{n-k} \) can be understood as a generalized function whose density is concentrated in a subspace \( \{ x \in \mathbb{R}^n \mid x_i = 0, i > k \} \) and equals \( f(x_{1:k}) \). It can be shown that the distribution acts on \( \phi \in \mathcal{S}(\mathbb{R}^n) \) in the following way:

\[
\langle T_f \otimes \delta^{n-k}, \phi \rangle = \int_{\mathbb{R}^k} f(x_{1:k})\phi(x_{1:k}, 0_{n-k}) dx_{1:k}
\]

Now to generalize the latter definition to any \( k \)-dimensional subspace we have to introduce a change of variables in tempered distributions.

Let \( g \in \mathcal{S}(\mathbb{R}^n) \) and \( U \in \mathbb{R}^{n \times n} \) be an orthogonal matrix, i.e. \( U^T U = I_n \). Then, \( g_U \in \mathcal{S}'(\mathbb{R}^n) \) is defined by the rule:

\[
\langle g_U, \phi \rangle = \langle g, \psi \rangle
\]

where \( \psi(x) = \phi(U^T x) \). If \( g = T_f \), the latter definition gives \( g_U = T f' \) where \( f'(x) = f(Ux) \).

Now, we define classes of tempered distributions:

\[
\mathcal{G}_k = \left\{ \langle f \otimes \delta^{n-k}, \phi \rangle \mid f \in \mathcal{S}(\mathbb{R}^k), U \in \mathcal{O}(n) \right\},
\]

and

\[
\mathcal{F}_k = \left\{ \langle T_f \otimes \delta^{n-k}, \phi \rangle \mid f \in \mathcal{S}(\mathbb{R}^k), U \in \mathcal{O}(n) \right\},
\]

where \( \mathcal{O}(n) = \{ U \in \mathbb{R}^{n \times n} \mid U^T U = I_n \} \). The first two classes are related as:

**Theorem 1.** \( \mathcal{G}_k = \mathcal{G}_k^* \).

The last two classes are isomorphic under the Fourier transform.

**Theorem 2.** \( \mathcal{F}[\mathcal{G}_k] = \mathcal{F}_k \) and \( \mathcal{F}^{-1}[\mathcal{F}_k] = \mathcal{G}_k \).

**Proof.** Let us prove first that if \( g = T_f \otimes \delta^{n-k} \), then \( \mathcal{F}[g] = T_r \), where \( r(x) = \tilde{f}(x_{1:k}), x \in \mathbb{R}^n \). For that we have to prove that \( \langle \mathcal{F}[g], \phi \rangle = \langle T_r, \phi \rangle \) for any \( \phi \in \mathcal{S}(\mathbb{R}^n) \). Indeed,

\[
\langle \mathcal{F}[g], \phi \rangle = \langle g, \mathcal{F}[\phi] \rangle = \langle T_f \otimes \delta^{n-k}, \int_{\mathbb{R}^n} \phi(y) e^{-ix^T y} dy \rangle =
\]

\[
\langle T_f, \int_{\mathbb{R}^n} \phi(y) e^{-ix_{1:k}^T y_{1:k}} dy \rangle = \int_{\mathbb{R}^n} f(x_{1:k}) \phi(y) e^{-ix_{1:k}^T y_{1:k}} dy dx_{1:k} = \int_{\mathbb{R}^n} \tilde{f}(y_{1:k}) \phi(y) dy = \langle T_r, \phi \rangle.
\]
Let us calculate the image of $G_k$ under the Fourier transform. It is easy to see that for any $g \in S'(\mathbb{R}^n)$, $\phi \in \mathcal{S}(\mathbb{R}^n)$ and orthogonal $U \in \mathbb{R}^{n \times n}$ we have:

\[
\langle \mathcal{F}[gu], \phi(x) \rangle = \langle gu, \mathcal{F}[\phi](x) \rangle = \langle g, \mathcal{F}[\phi](U^T x) \rangle = \\
\langle g, \mathcal{F}[(U^T x)] \rangle = \langle (\mathcal{F}[g])_U, \phi(x) \rangle.
\]

Therefore, $\mathcal{F}[gu] = (\mathcal{F}[g])_U$. Thus, if $g = T_f \otimes \delta^{n-k}$, then

\[
(\mathcal{F}[gu])_U = T_{r'}
\]

where $r'(x) = r(Ux) = \hat{f}(U_kx)$ and $U_k \in \mathbb{R}^{k \times n}$ is a matrix consisting of first $k$ rows of $U$. Thus, $T_{r'} \in \mathcal{F}_k$.

Let us show that by varying $f \in \mathcal{S}(\mathbb{R}^k)$ and $U$ in the expression $\hat{f}(U_kx)$ we can obtain any function from $\mathcal{F}_k$. For this, it is enough to show that $\mathcal{F}_k$ is equivalent to the following set of functions:

\[
\mathcal{Q} = \{g(U_kx) | g \in \mathcal{S}(\mathbb{R}^k), U_k \in \mathbb{R}^{k \times n}, U_kU_k^T = I_k\}
\]

The fact $\mathcal{Q} \subseteq \mathcal{F}_k$ is obvious. Let us now prove that $\mathcal{Q} \supseteq \{g(Px) | g \in \mathcal{S}(\mathbb{R}^k), P \in \mathbb{R}^{k \times n}, \text{rank} P = k\} = \mathcal{F}_k$. Indeed, if $f(x) = g(Px)$, then $f(x) = g'(U_kx)$ where $U_k = (PP^T)^{-1/2}P$ and $g'(y) = g((PP^T)^{1/2}y)$. By construction, $U_kU_k^T = I_k$ and $g' \in \mathcal{S}(\mathbb{R}^k)$. Thus, $\mathcal{Q} = \mathcal{F}_k$.

Therefore, $\mathcal{F}[G_k] = \mathcal{F}_k$, and from the bijectivity of the Fourier transform we obtain $\mathcal{F}^{-1}[\mathcal{F}_k] = G_k$.

For any collection $f_1, \ldots, f_l \in \mathcal{S}'(\mathbb{R}^n)$, $\text{span}_H\{f_i\}$ denotes $\{\sum_{i=1}^l \lambda_i f_i | \lambda_i \in \mathbb{R}\} \subseteq \mathcal{S}'(\mathbb{R}^n)$, which is a linear space over $\mathbb{R}$. The set $G_k'$ has the following simple characterization:

**Theorem 3.** For any $T \in \mathcal{S}'(\mathbb{R}^n)$, $T \in G_k'$ if and only if

\[
\dim \text{span}_\mathbb{R}\{x_1T, x_2T, \ldots, x_nT\} \leq k.
\]

Informally, the theorem holds because any linear dependency $\alpha_1 x_1 T + \cdots + \alpha_n x_n T = 0$ over $\mathbb{R}$ implies that if $\alpha_1 x_1 + \cdots + \alpha_n x_n \neq 0$, then $T = 0$. This is equivalent to a statement that the support of $T$ is concentrated on a subspace $\alpha_1 x_1 + \cdots + \alpha_n x_n = 0$. If $\dim \text{span}_\mathbb{R}\{x_1T, x_2T, \ldots, x_nT\} \leq k$, then one can find $n-k$ such dependencies, which means that the support of $T$ is $k$-dimensional.

Let $\mathcal{B}(\mathbb{R}^n)$ denote the Borel sigma-algebra on $\mathbb{R}^n$ and $\mathcal{P}$ denote a set of all Borel probability measures on $\mathbb{R}^n$. Let us now define

\[
\mathcal{P}_k = \{\mu \in \mathcal{P} | \forall \mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n, \forall A \in \mathcal{B}(\mathbb{R}^n): \mu(A) = \mu(A \cap \text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k))\}
\]

i.e. $\mathcal{P}_k$ is a set of probability measures with all probability concentrated in some subspace $\text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ whose dimension is not greater than $k$. It is easy to see that $T_{\mu} \in G_k'$ for any $\mu \in \mathcal{P}_k$.

**4 Examples of LDR formulations**

**Maximum mean discrepancy PCA (MMD-PCA)** Let $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a continuous Mercer kernel, and $\mathcal{H}_K$ be a reproducing kernel Hilbert space (RKHS) defined by $K$. The kernel $K(x, y)$ defines the so-called kernel embedding of probability measures $\phi$:

\[
\mu \in \mathcal{P} \xrightarrow{\phi} \mathbb{E}_{x \sim \mu} K(x, y) = \int K(x, y) d\mu(y) = \mu(\mathcal{X}).
\]

The Maximum Mean Discrepancy (MMD) distance [20] is defined as the distance induced by metrics on $\mathcal{H}_K$, i.e. for two probability measures $\mu, \nu \in \mathcal{P}$,

\[
d_{\text{MMD}}(\mu, \nu) = \|\phi(\mu) - \phi(\nu)\|_{\mathcal{H}_K}.
\]

Let $x_1, \ldots, x_N \in \mathbb{R}^n$ be the dataset of points. This dataset defines the empirical probabilistic measure $\mu_{\text{data}}$ that corresponds to the tempered distribution $T_{\mu_{\text{data}}} = \frac{1}{N} \sum_{i=1}^N \delta^n(x - x_i)$. We shall study a method concurrent to PCA that is based on solving the following problem:

\[
\min_{\nu \in \mathcal{P}_k} d_{\text{MMD}}(\mu_{\text{data}}, \nu) = \min_{\nu \in \mathcal{P}_k} \|\phi(\mu_{\text{data}}) - \phi(\nu)\|_{\mathcal{H}_K}
\]

i.e. we shall attempt to approximate the empirical probabilistic measure $\mu_{\text{data}}$ with another probabilistic measure $\nu$ which is supported in some $k$-dimensional subspace of $\mathbb{R}^n$. To our knowledge, the task [17] has not been yet considered in the research field of LDR.
Example 1 (Gaussian MMD-PCA). Let \( k(x) = G^n_h(x) \) where \( G^n_h(x) = \frac{e^{-\frac{|x|^2}{2h^2}}}{(2\pi h^n)^{n/2}} \) is the radial Gaussian kernel on \( \mathbb{R}^n \) and \( K(x, y) = (k \circ k)(x - y) = G^n_{2h}(x) \). For such a kernel, we have
\[
d_{\text{MMD}}(\mu, \nu) = \|\psi(\mu) - \psi(\nu)\|_{L_2(\mathbb{R}^n)},
\]
where \( \psi(\mu) = \int k(x - y)d\mu(y) \) is just a smoothing of the distribution \( \mu \) via the Weierstrass trasform.

In this example, as \( h \to +0 \), the optimal measure \( \nu^* = \arg \min_{\nu \in P_h} \|\psi(\mu_{\text{data}}) - \psi(\nu)\|_{L_2(\mathbb{R}^n)} \) is supported in a \( k \)-dimensional subspace that contains the largest possible number of points from \( \{x_1, \ldots, x_N\} \). Khachiyan demonstrated \([33]\) that the following problem is NP-hard: given \( \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^n \), find an \( n \) – 1-dimensional subspace of \( \mathbb{R}^n \) that contains at least \((1 - \varepsilon)(1 - \frac{1}{n})N \) points from the dataset. This indicates that in the regime \( h \to +0 \), the task \([17]\) is NP-hard. In other words, it is unlikely that the task admits an efficient algorithm, in general. In \([4]\) we describe an algorithm for the Gaussian MMD-PCA.

The higher moments PCA (HM-MMD-PCA) Another natural approach to measuring the similarity of two distributions is based on the difference between moments:
\[
d_{\text{HM}}(\mu, \nu)^2 = \sum_{s=1}^{4} \frac{\lambda_s}{n^s} \sum_{1 \leq i_1, \ldots, i_s \leq n} (m_{i_1, \ldots, i_s} - n_{i_1, \ldots, i_s})^2
\]
where \( m_{i_1, \ldots, i_s} = \mathbb{E}_{x \sim \mu}[X[i_1] \ldots X[i_s]] \) and \( n_{i_1, \ldots, i_s} = \mathbb{E}_{x \sim \nu}[X[i_1] \ldots X[i_s]] \) are corresponding moments. The positive parameters \( \lambda_1, \lambda_2, \lambda_3, \lambda_4 \) are chosen to fix the relative importance of the mean, the co-variance, the co-skewness and the co-kurtosis.

Thus, we will be interested in the following optimization task (analogous to \([17]\)):
\[
\min_{\nu \in P_h} d_{\text{HM}}(\mu_{\text{data}}, \nu)
\]
If we set \( \lambda_2 = 1 \) and \( \lambda_1 = \lambda_3 = \lambda_4 = 0 \), then the solution of the task \((21)\) coincides with the solution of the classical PCA. Let us briefly demonstrate that. Let \( X = [x_1, \ldots, x_N] \) be the data matrix, \( Y = [y_1, \ldots, y_N] \) be the SVD of \( X \) truncated at \( k \)-th term, \( \sigma_i(X) \) be an \( i \)-th singular value of \( X \). By \( \mu_{\text{pca}} \) we denote a probabilistic measure concentrated in points \( \{y_1\}_{i=1}^N \). In that case we have
\[
d_{\text{HM}}(\mu_{\text{data}}, \mu_{\text{pca}})^2 = \| \frac{1}{N} XX^T - \frac{1}{N} YY^T \|_F^2 = \frac{1}{N^2} \sum_{i=k+1}^{\min(N, n)} \sigma_i^4(X).
\]
But for any \( \nu \in P_h \) the covariance matrix \( \text{cov}(\nu) = [\mathbb{E}_{x \sim \nu} x_i x_j]_{i,j \in [n]} \) is of rank \( k \). Therefore, by Eckart-Young-Mirsky’s theorem, we have
\[
d_{\text{HM}}(\mu_{\text{data}}, \nu)^2 = \| \frac{1}{N} XX^T - \text{cov}(\nu) \|_F^2 \geq \frac{1}{N^2} \sum_{i=k+1}^{\min(N, n)} \sigma_i^4(X).
\]
Thus, the minimum of \( d_{\text{HM}}(\mu_{\text{data}}, \nu) \) is attained at \( \nu = \mu_{\text{pca}} \).

Thus, the task \((20)\) can be considered as a direct generalization of PCA that takes into account higher moments. Note that the distance based on higher moments is a special case of maximum mean discrepancy metric, where \( K(x, y) = \sum_{s=1}^{4} \frac{\lambda_s}{n^s} (x \cdot y)^s \). That is why we denote the task as HM-MMD-PCA. In Section \([7]\) we prove that there is an efficient 2-approximating algorithm for the HM-MMD-PCA. In \([1]\) we additionally describe another algorithm for the HM-MMD-PCA based on a generic alternating scheme.

Wasserstein distance PCA (WD-PCA) Another significant distance between probability measures with the origins in the transport theory is the Wasserstein distance (see \([34]\)).

Let \((\mathbb{R}^n, ||\cdot||)\) be a Banach space and \( p \geq 1 \). Between any two Borel probability measures \( \mu, \nu \) on \( \mathbb{R}^n \) with \( \int \|x\|^p d\mu < \infty \) and \( \int \|x\|^p d\nu < \infty \) the \( p \)th Wasserstein distance is:
\[
W_p(\mu, \nu) = \left( \inf_{\pi \in \Pi(\mu, \nu)} \int \|x - y\|^p d\pi \right)^{1/p}
\]
where \( \Pi(\mu, \nu) \) is a set of all couplings of \( \mu \) and \( \nu \). The Wasserstein distance defines another version of LDR problem:
\[
\min_{\nu \in P_h} W_p(\mu_{\text{data}}, \nu)
\]
In \([B]\) one can find proofs that in the case of \( l_1 \) norm \( \|x\| = \sum |x_i| \) and \( p = 1 \), the task \((22)\) corresponds to the well-studied robust PCA problem \([35]\). If, instead of the \( l_1 \)-norm, we use the \( l_2 \)-norm and set \( p = 1 \), this leads to another well-studied task, which is known as the outlier pursuit problem \([36, 37]\). In the case of the \( l_2 \)-norm and a general \( p \geq 1 \) we obtain the \( l_p \) subspace approximation problem \([38, 39]\). Note that, except for the \( l_2 \) subspace approximation problem, all these problems are NP-hard. In \([J]\) we describe an algorithm for the WD-PCA in the case of \( l_2 \)-norm and \( p = 1 \).
Sufficient dimension reduction with optimized regression function (SDR-ORF). Given a labeled dataset \( \{(x_i, y_i)\}_{i=1}^N \) where \( x_i \in \mathbb{R}^n, y_i \in C \) (\( C \) is a finite set of classes for a classification, or \( \mathbb{R} \) for a regression problem), the sufficient dimension reduction problem can be informally described as a problem of finding vectors \( w_1, \cdots, w_k \in \mathbb{R}^n \) such that conditional distributions satisfy \( p(y|w_1^T x, \cdots, w_k^T x) \approx p(y|x) \) (possibly, under some additional assumptions on the form of \( p(y|x) \)).

We formulate the SDR-ORF problem as an optimization task:

\[
\inf_{f \in \mathcal{F}_k} J(f)
\]

The object \( f : \mathbb{R}^n \to \mathbb{R} \) is a smooth real-valued function. We assume that \( f \) is a candidate for the regression function and \( J(f) \) is a cost function that values how strongly \( f \) fits in this role. In practice for the regression case and for the binary classification case with 0-1 outputs we use the following cost functions correspondingly:

\[
J(f) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\epsilon \sim N(0, \sigma^2 I_n)} |y_i - f(x_i + \epsilon)|^2
\]

and

\[
J(f) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\epsilon \sim N(0, \sigma^2 I_n)} H \left( y_i, \frac{e^{f(x_i + \epsilon)}}{1 + e^{f(x_i + \epsilon)}} \right)
\]

where \( H(y, p) = -y \log p - (1 - y) \log(1 - p) \) and \( \nu > 0 \) is a parameter.

By requiring \( f \in \mathcal{F}_k \), we assume that the regression function \( f \) satisfies (for \( k \) fixed in advance): \( f(x) = g(w_1^T x, \cdots, w_k^T x) \), where \( w_1, \cdots, w_k \in \mathbb{R}^n \). Thus, given an input \( x \), an output of \( f \) depends on the projection of \( x \) onto span\( (w_1, \cdots, w_k) \). The set span\( (w_1, \cdots, w_k) \) is called the effective subspace. In [2] we describe an algorithm for the SDR-ORF problem.

5 Reduction of the optimization problem to ordinary functions

The central problem that our paper addresses is the optimization of an objective function over \( \mathcal{G}'_k \)? In this section we suggest an approach based on penalty functions and kernels.

5.1 The definition of the penalty function

In this subsection we introduce a penalty function \( R(f) \). Let \( M : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{C} \) be some bounded function such that \( [M(z_i, z_j)]_{1 \leq i,j \leq n} \) is a positive semidefinite matrix for any \( \{z_i\}_{i \in [n]} \subseteq \mathbb{R}^n \), \( x \in \mathcal{X} \). For \( f, g : \mathbb{R}^n \to \mathbb{C} \) let us denote

\[
\langle f|M|g \rangle = \int_{\mathbb{R}^n \times \mathbb{R}^n} f(x)^* M(x, y) g(y) dx dy.
\]

For \( f, g \in L_1(\mathbb{R}^n) \),

\[
\langle f|M|g \rangle \leq \sup_{x,y} |M(x, y)| \cdot \|f\|_{L_1} \cdot \|g\|_{L_1} < \infty.
\]

For general \( f, g \in \mathcal{S}'(\mathbb{R}^n) \) the expression \( \langle f|M|g \rangle \) is defined if there are \( f_\epsilon, g_\epsilon \in L_1(\mathbb{R}^n) \) such that \( T_{f_\epsilon} = f \ast G^*_\epsilon \), \( T_{g_\epsilon} = g \ast G^*_\epsilon \) and \( \lim_{\epsilon \to 0} \langle f_\epsilon|M|g_\epsilon \rangle = A < \infty \). Then, \( \langle f|M|g \rangle \) defines \( A \). For example, for continuous \( M \) we have

\[
\langle \delta^n|M|\delta^n \rangle = M(0,0).
\]

One can build a Gram matrix from the collection of functions \( \{x_i f\}_{i=1}^n \), \( \{x_i f|M|x_j f\}_{1 \leq i,j \leq n} \). Let us denote a real part of the Gram matrix

\[
[(x_i f|M|x_j f)]_{1 \leq i,j \leq n}
\]

by \( M_f \).

Theorem 3 concludes, from \( f \in \mathcal{G}_k \), that \( \dim \text{span}_{\mathbb{R}} \{x_1 f, x_2 f, \cdots, x_n f\} \leq k \).

**Theorem 4.** Let \( M(x, y) \) be a bounded Lipschitz function. If \( f = (T_g \otimes \delta^{n-k})_U \in \mathcal{G}'_k \) is such that \( \{x_i g\}_{i=1}^k \subseteq L_1(\mathbb{R}^k) \), then \( \langle x_i f|M|x_j f \rangle \) is defined and \( \text{rank } M_f \leq k \).

---

1Throughout the paper the kernel that induces the MMD distance is denoted by \( K \) and the kernel that is used to define a penalty is denoted by \( M \).
Definition 1. Let $A \in \mathbb{R}^{n \times n}$ be a positive semidefinite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ (with counting multiplicities). Then, the Ky Fan $k$–anti-norm of $A$ is $\|A\|_k = \sum_{i=1}^{k} \lambda_{n+1-k}$.

Let

$$R(f) = \|M_f\|_{n-k}.$$  

By construction, by penalizing the value of $R(f)$, we enforce $M_f$ to be close to some matrix of rank $k$. Equivalently, we enforce a real part of the Gram matrix of $\{x_i, f\}_{i \in [n]}$ to be of close to a rank $k$ matrix. By Theorem 3, the condition $\dim \text{span}_R \{x_1, x_2, \cdots, x_n, f\} \leq k$ implies $f \in G_k^\perp$, therefore, we enforce $f$ to be close to some function from $G_k^\perp$.

In the next section we will justify the latter informal logic by reducing the optimization over ordinary functions with the penalty function $\inf R_M$ to the optimization over ordinary functions with the penalty function $R(f)$.

5.2 Proper kernels

For a function $M(x, y) : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{C}$, let us denote by $O_M$ a linear operator between $\text{Dom}(O_M)$ and $L_2(\mathbb{R}^n)$ given by $O_M[f] = \int_{\mathbb{R}^n} M(x, y)f(y)dy$ where $\text{Dom}(O_M) = \{f \in L_2(\mathbb{R}^n) | O_M[f] \in L_2(\mathbb{R}^n)\}$. For any operator $O$ between spaces $H_1$ and $H_2$, we denote its range by $\text{Range}[O] = \{O(x)|x \in H_1\}$.

**Definition 2.** The function $M(x, y) : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{C}$ is called the proper kernel if and only if

1. $O_M : L_2(\mathbb{R}^n) \rightarrow L_2(\mathbb{R}^n)$ is a properly defined, strictly positive and self-adjoint operator,

2. $\max_{x, y} |M(x, y)| < \infty$,

3. $\text{Range}[O_M] \cap S(\mathbb{R}^n) = S(\mathbb{R}^n)$.

Note that the latter definition implies that $M(y, x) = M(x, y)^*$ (modulo some null set) and $\langle f, O_M[f] \rangle_{L_2(\mathbb{R}^n)} > 0, \forall f \in L_2(\mathbb{R}^n)$, $f \neq 0$.

**Example 2.** The Gaussian kernel is of special interest in applications: $M(x, y) = G^\perp_n(x - y)$.

It is captured by the following lemma:

**Lemma 1.** If $\zeta, \hat{\zeta} \in C(\mathbb{R}^n)$ are bounded, $\forall x \zeta(x) > 0$, then $M(x, y) = \zeta(x - y)$ is a proper kernel.

**Proof.** Verification of the first three conditions is easy, so we only check the fourth condition. Let us denote linear operators $C_\zeta[f] = \zeta \ast f$ and $O_\zeta[f](x) = g(x) f(x)$. Then we have $\mathcal{F}[C_\zeta[L_2(\mathbb{R}^n)]] = O_\zeta[L_2(\mathbb{R}^n)] \supseteq C_c^\infty(\mathbb{R}^n)$. Therefore, $\text{Range}[O_M] = C_\zeta[L_2(\mathbb{R}^n)] \supseteq F^{-1}[C_c^\infty(\mathbb{R}^n)]$. Since $C_c^\infty(\mathbb{R}^n)$ is dense in $S(\mathbb{R}^n)$, then $F^{-1}[C_c^\infty(\mathbb{R}^n)]$ also has this property. Thus, $\text{Range}[O_M] \cap S(\mathbb{R}^n) = S(\mathbb{R}^n)$.

Besides the Gaussian kernel the lemma also captures a case of the Laplace kernel $\zeta(x) = e^{-|x|}$. It is well-known that the Fourier transform of the Laplace kernel is the Poisson kernel: $\hat{\zeta}(x) = \frac{c_n}{(1 + |x|^2)^\frac{n}{2}}$ (which is also proper).

For $I : G_k^\perp \cup S(\mathbb{R}^n) \rightarrow \mathbb{R}^+,$ it is natural to reduce the optimization task over tempered distributions

$$I(f) \rightarrow \min_{f \in G_k^\perp} I(f)$$  

(29)

to an optimization task over ordinary functions with a penalty term $R$,

$$I(f) + \lambda \|M_f\|_{n-k} = I(f) + \lambda R(f) \rightarrow \inf_{f \in \mathcal{G}} I(f) + \lambda R(f) \rightarrow +0 \text{ for } \lambda_n \rightarrow +\infty \text{ (i.e. } \{f_n\} \text{ solves (30) for arbitrarily large values of the regularization parameter), then there exists a growing subsequence } \{n_k\} \text{ such that } T f_{n_k} \rightarrow^* T \text{ (weakly) where } T \text{ is a solution of (29)}.$$  

We make a thorough theoretical analysis of the case $\mathcal{G} = S(\mathbb{R}^n)$. If to formulate in a simplified way, for the last property to hold, the sequence $\text{Tr}(M_{f_n})$ should be bounded. Details on the conditions under which this reduction holds can be found in the following subsection.
5.3 Regular solutions and reduction theorems for $\mathfrak{F} = S(\mathbb{R}^n)$

For a sequence $\{f_i\}_{i=1}^{\infty} \subseteq S'(\mathbb{R}^n)$, $\lim_{s \to \infty} f_s$ denotes a set of points $f \in S'(\mathbb{R}^n)$, such that there exists a growing sequence $\{s_i\} \subseteq \mathbb{N}$ and $\lim_{i \to \infty} f_{s_i} = f$.

For $I : \mathcal{G}_k' \cup S(\mathbb{R}^n) \to \mathbb{R}^+$, it is natural to reduce the optimization task (29) to an optimization task over ordinary functions with a penalty term (30). To have an equivalence between (29) and (30) we need to assume that $I$’s behaviour when approaching $f \in \mathcal{G}_k'$ from a set $S(\mathbb{R}^n)$ is continuous, i.e. for any sequence $\{f_i\} \subseteq S(\mathbb{R}^n)$ such that $T_{f_i} \to f \in \mathcal{G}_k'$, we have $\lim_{i \to \infty} I(T_{f_i}) = I(f)$.

Let us introduce the notion of a regular solution both for (29) and (30). Let

$$B_k = \bigcup_{C > 0} \{f \in \mathcal{G}_k | \text{Tr}(M_f) \leq C\}.$$  \hspace{1cm} (31)

**Definition 3.** Any $f \in \text{Arg min}_{f \in \mathcal{G}_k} I(f) \cap B_k$ is called a regular solution of (29).

In other words, $B_k$ formalizes a set of distributions from $\mathcal{G}_k'$, that can be approached through sequences $\{f_i\} \subseteq \mathcal{G}_k$ for which $\text{Tr}(M_{f_i})$ does not blow up. Obviously, $\mathcal{G}_k \subseteq B_k \subseteq \mathcal{G}_k'$. In applications, regular solutions include all $\text{Arg min}_{f \in \mathcal{G}_k} I(f)$ if we choose the kernel $M$ correctly. This regularity is important for a reduction to the penalty form (30), because when approaching a non-regular solution we are unable to guarantee a bounded behaviour of $M_f$ (and of $R(f)$).

**Definition 4.** A sequence $\{f_i\}_{i=1}^{\infty} \subseteq S(\mathbb{R}^n)$ is said to solve (30) if

$$I(f_i) + \lambda i R(f_i) \leq \inf_{f \in S(\mathbb{R}^n)} I(f) + \lambda i R(f) + \epsilon_i$$  \hspace{1cm} (32)

where $\epsilon_i \to +0$ and $\lambda i \to +\infty$, $i \to +\infty$. If, additionally, $\text{Tr}(M_{f_i})$ is bounded, then $\{f_i\}_{i=1}^{\infty}$ is said to solve (30) regularly.

Let us define

$$\text{rsol } (I(f), R(f)) = \lim_{\{f_i\}_{i=1}^{\infty} \text{ r.solves } (11)} \text{Tr}(M_{f_i}) \hspace{1cm} (33)$$

**Theorem 5.** If $M$ is a proper kernel, then $\text{rsol } (I(f), R(f)) \subseteq \text{Arg min}_{f \in \mathcal{G}_k} I(f)$.

**Theorem 6.** If $M$ is a proper kernel and $\text{rsol } (I(f), R(f)) \neq \emptyset$, then

$$\text{Arg min}_{f \in \mathcal{G}_k} I(f) \cap B_k \subseteq \text{rsol } (I(f), R(f))$$

**Theorem 7** (Reduction theorem). If $M$ is a proper kernel, $\text{Arg min}_{f \in \mathcal{G}_k} I(f) \subseteq B_k$ and $\text{rsol } (I(f), R(f)) \neq \emptyset$, then $\text{rsol } (I(f), R(f)) = \text{Arg min}_{f \in \mathcal{G}_k} I(f)$.

Suppose that we now solve a sequence of problems (30) and find $\{f_s\}_{s=1}^{\infty}$. According to Theorems 5 and 6 the following are potential scenarios:

1. $\text{Tr}(M_{f_s})$ blows up and the convergence is not guaranteed. This situation can be avoided by controlling $\text{Tr}(M_f)$ in an optimization process. In practice, when $f$ has a parameterized form, this can be done by bounding parameters. If $\text{Tr}(M_{f_s})$ does not blow up, we still have two subcases:

   2.1 $\lim_{s \to \infty} T_{f_s} \neq \emptyset$. This implies a positive outcome to approach (30) to the optimization problem, Problem (29).

   2.2 $\lim_{s \to \infty} T_{f_s} = \emptyset$. This exotic situation can happen only if a sequence $T_{f_s}$ leaves any sequentially compact subset of $S'(\mathbb{R}^n)$. Bounding parameters also tackles this case.

Let us now concentrate on the task (30) and describe the alternating scheme for its solution.

6 The alternating scheme

We will concentrate on problem (30). It is known [25] that the Ky Fan anti-norm is a concave function, i.e. $R(\phi) = \|M_\phi\|_{n \to k}$ depends on $M_\phi$ in a concave way. It can be shown that the dependence of $R(\phi)$ on $\phi$ is both non-convex and non-concave, i.e. we deal with a non-convex optimization task.
Let $B(H_1, H_2)$ denote a set of bounded linear operators between Hilbert spaces $H_1$ and $H_2$. For $O \in B(H_1, H_2)$ the rank of $O$ is defined as $\dim \mathcal{R}(O)$. Let $L_2^0(R^n)$ be the Hilbert space (over $R$) of real-valued functions from $L_2(R^n)$ (i.e. the real-valued $L_2$-space) and $L_2^0(R^n) = L_2^0(R^n) \times L_2^0(R^n)$). The space $L_2^0(R^n)$ is equivalent to $L_2(R^n)$ treated as a linear space over $R$. Below we do not distinguish $[\phi_1, \phi_2] \in L_2^0(R^n)$ and $\phi_1 + i\phi_2 \in L_2(R^n)$. It is easy to see that any $O \in B(L_2^0(R^n), R^n)$ can be given by formula:

$$O[\phi]_i = \text{Re} \langle O_i, \phi \rangle_{L_2^0(R^n)}, O_i \in L_2(R^n), i = 1, n,$$

i.e. $O \in B(L_2^0(R^n), R^n)$ can be identified with a vector of functions $O = [O_i]_{i=1,n}, O_i \in L_2(R^n)$ and the Hilbert–Schmidt norm on $B(L_2^0(R^n), R^n)$ (i.e. $\sqrt{\text{Tr} O^*O}$) is

$$\|O\| = \sqrt{\sum_{i=1}^n \|O_i\|^2_{L_2^0(R^n)}}.$$  

(34)

Recall that for a Mercer kernel $M$, $O_M[\phi](x) = \int_{R^n} M(x, y)\phi(y)dy$ is a positive operator whose domain is $\text{Dom}(O_M) = \{f \in L_2(R^n) \mid O_M[f] \in L_2(R^n)\}$ and range is a subset of $L_2(R^n)$. If we assume that $\text{Dom}(O_M)$ is dense in $L_2(R^n)$, then its adjoint $O_M^*$ and the square root $\sqrt{O_M} : \text{Dom}(O_M) \to L_2(R^n)$ can be properly defined [40].

Thus, $O_M$ is self-adjoint. For any complex-valued function $f$ such that $\text{Tr} Mf < \infty$ let us introduce a linear operator $S_f : L_2^0(R^n) \to R^n$ by the following rule:

$$S_f[\phi]_i = \text{Re} \langle \sqrt{O_M[x_i f(x)]}, \phi \rangle_{L_2^0(R^n)},$$

i.e. $(S_f)_i = \sqrt{O_M[x_i f(x)]}$, $i = 1, n$. In the latter definition the expression $\langle \sqrt{O_M[x_i f(x)]}, \sqrt{O_M[x_i f(x)]}\rangle_{L_2^0(R^n)} = (M_f)_{ii} < \infty$ and the Cauchy-Schwarz inequality.

**Theorem 8.** Let $M$ be a Mercer kernel such that $\text{Dom}(O_M)$ is dense in $L_2(R^n)$ and $\text{Tr} Mf < \infty$. Then, $S_f \in B(L_2^0(R^n), R^n)$ and $S_f S_f^* = M_f$. Moreover,

$$R(f) = \min_{S \in B(L_2^0(R^n), R^n), \text{rank }S \leq k} \|S - S_f\|_2^2$$

and the minimum is attained at $S = P_f S_f$ where $P_f = \sum_{i=1}^k u_i u_i^*$ and $\{u_i\}_{i=1}^k$ are unit eigenvectors of $M_f$ corresponding to the $k$ largest eigenvalues (counting multiplicities).

**Proof.** The boundedness of $S_f$ follows from the Cauchy-Schwarz inequality:

$$|S_f[\phi]_i|^2 = |\text{Re} \langle \sqrt{O_M[x_i f(x)]}, \phi \rangle|^2 \leq \langle \sqrt{O_M[x_i f(x)]}, \sqrt{O_M[x_i f(x)]}\rangle \langle \phi, \phi \rangle = \langle x_i f, O_M[x_i f]\rangle \langle \phi, \phi \rangle$$

and therefore:

$$\|S_f[\phi]\|^2 = \sum_{i=1}^n |S_f[\phi]_i|^2 \leq \text{Tr} Mf \|\phi\|_{L_2^0(R^n)}^2.$$  

(39)

Thus, we have checked that $S_f$ is bounded.

By definition, $S_f^T : R^n \to L_2^0(R^n) \times L_2^0(R^n)$ and $\{u, S_f[\phi_1, \phi_2] = (S_f^T[u], [\phi_1, \phi_2]), u \in R^n, [\phi_1, \phi_2] \in L_2^0(R^n) \times L_2^0(R^n)$). Let us denote $f_1 = \text{Re} f, f_2 = \text{Im} f$. It is easy to see that the following operator satisfies the latter identity:

$$O[u] = [\sqrt{O_M[f_1(x)x^T u]}, \sqrt{O_M[f_2(x)x^T u]}].$$  

(40)

Since the adjoint is unique, then $S_f^T = O$. Let us calculate $S_f S_f^T$:

$$u \mapsto [\langle x_1 f_1(x), \sqrt{O_M[f_1(x)x^T u]} \rangle, \ldots, \langle x_n f_1(x), \sqrt{O_M[f_1(x)x^T u]} \rangle] + [\langle x_1 f_2(x), \sqrt{O_M[f_2(x)x^T u]} \rangle, \ldots, \langle x_n f_2(x), \sqrt{O_M[f_2(x)x^T u]} \rangle]$$

$$= [\text{Re} \langle x_i f, M[x_j f]\rangle]_{1 \leq i, j \leq n} u = M_f u$$

(41)
Thus, \( S_f S_f^\dagger = M_f \). Since \( \text{Tr} S_f S_f^\dagger < \infty \) and \( \| S_f^\dagger [u] \|^2 \leq \langle u, M_f u \rangle \), we obtain \( S_f^\dagger \) is a bounded operator.

Let \( u_1, \ldots, u_n \) be orthonormal eigenvectors of \( M_f = S_f S_f^\dagger \) and \( \lambda_1 \geq \cdots \geq \lambda_n > 0 \) be corresponding nonzero eigenvalues. For \( \sigma_i = \sqrt{\lambda_i} \) let us define \( v_i = \frac{S_f^\dagger [u_i]}{\sigma_i} \). Vector \( v_i \) corresponds to a pair of functions

\[
\phi_i = \frac{1}{\sigma_i} \left[ \sqrt{\mathcal{O}_{1\!\!f}[f_2(x)x^T_u_i]} \right] \in L_2^r(\mathbb{R}^n) \times L_2^r(\mathbb{R}^n)
\]

(42)

It is easy to see that \( v_1, \ldots, v_n \) is an orthonormal basis in \( \text{Im} S_f^\dagger \), and \( S_f^\dagger \) can be expanded in the following way:

\[
S_f^\dagger = \sum_{i=1}^{n'} \sigma_i v_i^\dagger u_i^\dagger,
\]

(43)

and therefore, SVD for \( S_f \) is

\[
S_f = \sum_{i=1}^{n'} \sigma_i u_i v_i^\dagger.
\]

(44)

By the Eckart-Young-Mirsky theorem (see Theorem 4.4.7 from [41]), an optimal \( S \) in

\[
\min_{S \in \mathcal{B}(L_2^r(\mathbb{R}^n), \mathbb{R}^n), \text{rank} S \leq k} \| S_f - S \|^2_*
\]

is defined by a truncation of SVD for \( S_f \) at \( k \)th term, i.e.

\[
S = \sum_{i=1}^k \sigma_i u_i v_i^\dagger = P_f S_f,
\]

(45)

where \( P_f = \sum_{i=1}^k u_i u_i^\dagger \) is a projection operator to first \( k \) principal components of \( M_f \). Moreover, \( \| S_f - P_f S_f \|^2 = \sum_{i=k+1}^{n'} \sigma_i^2 = \| M_f \|_{n-k} = R(f) \).

Thus, it is natural to view the Task (30) as a minimization of \( I(\phi) + \lambda \| S_\phi - S \|^2_* \) over two objects: \( f \in \mathfrak{F} \) and \( S \in \mathcal{B}(L_2^r(\mathbb{R}^n), \mathbb{R}^n) : \text{rank} S \leq k \). The simplest approach to minimize a function over two arguments is to optimize alternatingly, i.e. first over \( f \), and then over \( S : \text{rank} S \leq k \), and so on. Theorem 8 gives that the minimization over \( S \) is equivalent to the truncation of SVD(\( S_f \)) at the \( k \)-th term. This idea, that we dub the alternating scheme (AS), is described in Algorithm 1.

**Algorithm 1** The alternating scheme (AS) for (30)

\[
(P_0,S_{\phi_0}) \leftarrow \text{Initialize for } t = 1, \ldots, T \text{ do }
\phi_t \leftarrow \text{arg min } \phi \in \mathfrak{F} \min I(\phi) + \lambda \| S_\phi - P_{t-1} S_{\phi_{t-1}} \|^2_* \text{ (minimizing over } \phi) \text{ Calculate } M_{\phi_t} \text{ and find } \{v_i\}_{i=1}^n \text{ s.t. } M_{\phi_t} v_i = \lambda_i v_i, \lambda_1 \geq \cdots \geq \lambda_n \text{ s.t. } P_t \leftarrow \sum_{i=1}^k v_i v_i^\dagger \text{ (Truncated SVD}(S_{\phi_t}) \text{ is } P_t S_{\phi_t})
\]

Output: \( v_1, \ldots, v_k \)

The alternating algorithm [1] allows for a reformulation in the dual space. By this we mean that in Algorithm 1 we substitute \( \phi_t \) for the original \( \phi_t \). If the primal Algorithm [1] deals with operators \( S_{\phi_t}^\dagger S_{\phi_{t-1}}^\dagger \), the dual version deals with vectors of functions \( \sqrt{G_{\sigma}} \frac{\partial \phi}{\partial x} \), \( \sqrt{G_{\sigma}} \frac{\partial \phi_{t-1}}{\partial x} \). Details of the dual algorithm can be found in [4].

The objective \( I(f) + \lambda R(f) \) can have many local minima due to the effect of the penalty term \( R(f) \). Therefore, the Alternating Scheme [1] is strongly dependant on the initialization step. One of such initialization procedures for the task [17] is described in the next section.
An approximate algorithm for the MMD-PCA

Let us apply Theorem 8 to the kernel $K(x, y) = (x \cdot y)H(x, y)$ and $H$ is a Mercer kernel (by construction, $K$ is also a Mercer kernel). In this section we demonstrate that, given a distribution $f(x) = \frac{1}{N} \sum_{i=1}^{N} \delta^n(x - x_i)$, a good guess for a $k$-dimensional space in which an optimal solution is supported is a span of the first $k$ principal components of $H_f$ (see the Algorithm 2).

**Algorithm 2** An approximate algorithm for (17) where $K(x, y) = (x \cdot y)H(x, y)$

| Input: $x_1, \ldots, x_N$, $f(x) = \frac{1}{N} \sum_{i=1}^{N} \delta^n(x - x_i)$ | Output: $x'_i = P x_i, i \in [N], f'(x) = \frac{1}{N} \sum_{i=1}^{N} \delta^n(x - x'_i)$ |
|---|---|
| Calculate $H_f = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j^T H(x_i, x_j)$ | |
| Calculate SVD($H_f$): $H_f = \sum_{i=1}^{n} \lambda_i v_i v_i^T$ where $\lambda_1 \geq \cdots \geq \lambda_n$ | |
| $P \leftarrow \sum_{i=1}^{k} v_i v_i^T$ | |

A specifics of this type of kernels is that the MMD distance (induced by $K$) till an optimal solution of the MMD-PCA task is bounded below by the Ky Fan $k$-antinorm of $H_f$, as shown in the following theorem.

**Theorem 9.** Let $K(x, y) = (x \cdot y)H(x, y)$ where $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a Mercer kernel, Dom($O_K$) is dense in $L_2(\mathbb{R}^n)$ and $f$ is such that $Tr H_f < \infty$. Then,

$$\inf_{f' \in G_k} \| f - f' \|^2_K \geq \sum_{i=k+1}^{n} \lambda_i$$

(47) where $\| g \|^2_K = \langle g | K | g \rangle$, $\lambda_1 \geq \cdots \geq \lambda_n$ are eigenvalues (counting multiplicities) of $H_f$.

**Sketch.** Let us apply Theorem 8 to the kernel $H$ and the function $f$. Recall that an element $O \in B(L_2(\mathbb{R}^n), \mathbb{R}^n)$ can be identified with a vector-function $[O_i]_{i=1}^{n}$, $O_i \in L_2(\mathbb{R}^n)$ where $O|\phi\rangle_i = \text{Re} \langle O_i \phi \rangle_{L_2(\mathbb{R}^n)}$. Since $S_f$ corresponds to $[\sqrt{O_H[x_i f(x)]}]_{i=1}^{n}$, the representation of Theorem 8 gives us

$$\| H_f \|_{n-k} = \min_{S \in B(L_2(\mathbb{R}^n), \mathbb{R}^n), \text{rank} S \leq k} \sum_{i=1}^{n} \| \sqrt{O_H[x_i f(x)]} - S_i(x) \|^2_{L_2(\mathbb{R}^n)}$$

(48)

The restriction $S \leq k$ is equivalent to $\text{dim} L_S \leq k$ where $L_S = \{ \sum_{i=1}^{n} \xi_i S_i(x) \ | \ \xi_i \in \mathbb{R} \}$ and $S$ corresponds to $[S_i]_{i=1}^{n}$.

Let $f' \in G_k$. By Theorem 8 we have $\text{dim} \text{span}_k \{ x_1 f', \ldots, x_n f' \} \leq k$. By Theorem 4, $\langle x_i f' | H | x_i f' \rangle$ is finite, therefore $\sqrt{O_H[x_i f'(x)]}$ can be properly defined and is in $L_2(\mathbb{R}^n)$. Therefore,

$$\text{dim} \text{span}_k \{ \sqrt{O_H[x_1 f'(x)]}, \ldots, \sqrt{O_H[x_n f'(x)]} \} \leq k$$

(49)

For any $f' \in G_k$ one can set $S_i = \sqrt{O_H[x_i f'(x)]}$ and search over all possible $f' \in G_k$ in the minimization operator. Thus,

$$R(f) \leq \inf_{f' \in G_k} \sum_{i=1}^{n} \| \sqrt{O_H[x_i f(x)]} - \sqrt{O_H[x_i f'(x)]} \|^2_{L_2(\mathbb{R}^n)} = \inf_{f' \in G_k} \| f - f' \|^2_K.$$  

(50)

Let $\mu_{\text{data}}$ be a uniform distribution over $\{x_i\}_{i=1}^{N}$ and $d_{\text{MMD}}$ be the MMD distance induced by $K(x, y) = (x \cdot y)H(x, y)$. The last theorem can be applied to a smoothed empirical distribution $f_\varepsilon(x) = \frac{1}{N} \sum_{i=1}^{N} G^\varepsilon(x - x_i)$ and then, we can send $\varepsilon \to 0$. All the more, the inequality will be satisfied if we search over $\mu \in \mathcal{P}_k$ due to $T_\mu \in G_k$.

Thus,

$$\inf_{\mu \in \mathcal{P}_k} d_{\text{MMD}}(\mu_{\text{data}}, \mu)^2 \geq \lim_{\varepsilon \to 0} \inf_{f' \in G_k} \| f_\varepsilon - f' \|^2_{L_2(\mathbb{R}^n)} = \inf_{f' \in G_k} \| f - f' \|^2_K$$

(51)

where $\lambda_1 \geq \cdots \geq \lambda_n$ are eigenvalues (counting multiplicities) of $H_f$, $f(x) = \frac{1}{N} \sum_{i=1}^{N} \delta^n(x - x_i)$. Thus, ($\sum_{i=k+1}^{n} \lambda_i)^{1/2}$ is a lower bound of the solution of (17).

For such an important practical case as the HM-MMD-PCA, a multiple of the square root of the Ky Fan $k$-antinorm of $H_f$ is also an upper bound.
Theorem 10. Let $H(x, y) = P(x \cdot y)$ where $P(x) = c_0 + c_1 x + \cdots + c_{l-1} x^{l-1}$, $c_i \geq 0$, $i \in [l-1]$, $f(x) = \frac{1}{N} \sum_{i=1}^{N} \delta^0(x - x_i) = T_{\mu_{\text{data}}}$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are eigenvalues of $H_f$. Then,
\[
\inf_{\mu \in P_k} d_{\text{MMD}}(\mu_{\text{data}}, \mu) \leq \sqrt{l} \left( \sum_{i=k+1}^{n} \lambda_i \right)^{1/2}. \tag{52}
\]

The following corollary is straightforward from the last theorem.

Corollary 1. A 2-approximating solution of the task (20) can be efficiently found by the Algorithm 2.

For the case when $H$ is the Gaussian kernel the situation is slightly trickier.

Theorem 11. Let $H(x, y) = e^{-\frac{|x-y|^2}{2\sigma^2}}$, $m_i = \int_{\mathbb{R}^n} ||x||^2 f(x) dx < \infty$, $i \in [4]$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are eigenvalues of $H_f$. Then,
\[
\inf_{f' \in \mathcal{U}_k} \|f - f'\|^2_{K} \leq M \sum_{j=k+1}^{n} \lambda_j^{1/2}, \tag{53}
\]

where $M = O(m_2 + \sqrt{2} \sigma \sqrt{m_4 m_2} + \sqrt{2} \sigma m_3)$.

An analogous theorem can be proved for $H(x, y) = (1 + \sigma^2 \|x - y\|^2)^{-\frac{n+1}{2}}$, i.e. the Poisson kernel.

8 Experiments

The alternating scheme \cite{[1]} is a general optimization method that needs to be specified for every optimization task. We designed numerical specifications of the alternating scheme \cite{[1]} for all 4 optimization tasks: (17), (20), (22) and (23) and made experiments with all of them. Details of the algorithms, i.e. numerical methods to minimize over $\phi$ and calculate $M_{\phi,n}$, can be found in Appendix. Note that for WD-PCA \cite{[22]}, we exploit the alternating scheme in the initial form (i.e. \cite{[1]}), and for MMD-PCA \cite{[17]}, HM-MMD-PCA \cite{[20]} and SDR-ORF \cite{[23]} we use the dual version of the scheme.

Behaviour of the Gaussian MMD-PCA for small $h$. We studied the difference in the behavior of PCA and a solution of (17), for the distance function induced by the kernel $K(x, y) = \frac{1}{\sqrt{(8\pi \sigma^2)^n}} e^{-\frac{||x||^2}{2\sigma^2}}$, obtained by the alternating scheme \cite{[1]} (AS for MMD-PCA), for the case when $h$ is small compared to the standard deviation of features. Experiments show that they are sharply different when data points are sampled along a low-dimensional manifold $\mathcal{M}$, which is bent globally, goes through the origin $O$ and has a large curvature at $O$ (see Fig. 1a). Since generated points do not lie on an affine subspace, the global nature of PCA makes it hard to interpret principal directions.

We select a smooth function $f: \mathbb{R}^{n-1} \to \mathbb{R}$, such that $f(0) = 0$ and generate points in the following way: points $x_1, x_2, \cdots, x_N \sim [-10, 10]^{n-1}$ are sampled uniformly, after calculation of $y_i = f(x_i)$ we add some noise: $z_i = (x_i, y_i) + \epsilon_i, \epsilon_i \sim N(0, 0.01 I_n)$. Both PCA and MMD-PCA are applied to the dataset (first 3 pictures on Figure 1a). As we see, MMD-PCA, unlike PCA, tries to catch ideal alignments of points rather than searching for a global alignment of points (which is non-existent). This property of MMD-PCA makes it a promising tool for a calculation of the tangent space to a data manifold at a given point. Fourth picture shows that when we have 2 equally important directions in data such that the principal direction of PCA is between them (red line), and we set $k = 1$, then MMD-PCA (green line) always chooses one of those directions. These experimental results are aligned with the theoretical observation given in Example \cite{[1]} in which we show that the Gaussian MMD-PCA task for $h \to 0^+$ is equivalent to finding a $k$-dimensional subspace that contains as many points of a dataset as possible. Thus, the Gaussian MMD-PCA can be considered as a method that can be potentially used to tackle the latter NP-hard problem. Some informal discussion of this problem can be found in \cite{[42]}.

Experiments with outlier detection (MMD-PCA, HM-MMD-PCA, WD-PCA). Following the experiment setup of \cite{[17]}, we choose parameters $N = n = 400, \delta = 0.05$ (or 0.1), $k = 10$ and generate random matrices $A \in \mathbb{R}^{N(1-\delta) \times k}, B \in \mathbb{R}^{n \times k}$ whose entries are iid as $N(0, 1)$. Then, columns of the matrix $BA^T \in \mathbb{R}^{n \times N(1-\delta)}$ (whose rank is $\leq k$) are concatenated with columns of the matrix $C \in \mathbb{R}^{n \times N\delta}$. $X = \text{concat}(BA^T, C) \in \mathbb{R}^{n \times N}$. The entries in $C$ are either iid as $N(0, 1)$ (case I) or $N\delta$ copies of the same vector whose entries are iid as $N(0, 1)$ (case II). Let $X = [x_1, \cdots, x_N]$, i.e. columns of $X$ are the data points. Thus, $N(1-\delta)$ columns of $BA^T$ lie in a $k$-dimensional subspace of $\mathbb{R}^n$ and $N\delta$ columns of $C$ are outliers, and solutions of tasks \cite{[17], [20]} for this dataset are expected to be supported in a column space of $BA^T$. 

14
After every iteration (step $t$ of the alternating scheme) we calculate the Frobenius distance between the projection operator $P_t$ of $I$ and the projection operator $P$ to the column space of $B A^T$, i.e., $\|P_t - P\|_F$. For the task (22), the dependence of $\|P_t - P\|_F$ on $t$ for different values of parameters $\delta$ and $\lambda$ is shown in Figure 1b. For tasks (17), (20) the behaviour of the alternating scheme is similar, 7 iterations are enough to approach the optimal subspace.

One of main goals of this experimental setup was to study how the kernel $M$, that defines the regularizer $R(f)$ by equation (28), affects the quality of a solution. Besides the speed of convergence we were interested in how $\|P^* - P\|_F$, where $P^* = \lim_{t \to \infty} P_t$ is the final projection operator (e.g. $P_{20}$ in practice), depends on the parameter $\sigma$ of the kernel $M = G_{\sigma}^T$ (bandwidth). It is natural to expect the quality of the solution $P^*$ to degrade as $\sigma \to +\infty$ (this corresponds to $M(x, y) \to 0$), and, less trivially, as $\sigma \to 0$ (this corresponds to $M(x, y) \to \delta^n(x - y)$). As the right plot on Figure 1b shows, for the HM-MMD-PCA, the solution $P^*$ is close to the correct $P$ if the bandwidth $\sigma$ is in interval $[e^{-2}, e^2]$ and it degrades beyond that interval. For the Gaussian MMD-PCA the degrading occurs beyond $[e^{1.3}, e^3]$. For the WD-PCA the interval for $\sigma$ is slightly narrower than $[e^{1.3}, e^3]$. Our numerical specification of the alternating scheme for WD-PCA involves training regularized Generative Adversarial Network (see for details) and are based on numerically unstable algorithms for the Wasserstein distance minimization. Finding numerical specifications for WD-PCA with a more stable behavior is a future work.

**Experiments with SDR-ORF.** We made experiments on the standard datasets, *Heart*, *Breast Cancer*, *Ionosphere*, *Diabetes*, *Boston House Prices* and *Wine Quality*. First, we applied the Sliced Inverse Regression algorithm (SIR) [13] to the training set and calculated the effective subspace for $k = 2, 3$. All points were projected onto that space and we obtained two- or three-dimensional representations of input points. In the last step we applied the ten nearest neighbors algorithm (KNN) to predict outputs (based on reduced inputs) on the test set (for the regression case, the 10-KNN regression was used). The same scheme was repeated with PCA, Kernel Dimensionality Reduction (KDR) algorithm [18] and the alternating scheme [1] (AS) adapted for the SDR-ORF.

We experimented with the dual version of algorithm [1] setting (after the data was standardized) the kernel’s parameter $\sigma = 0.5^2$ and $\lambda = 10.0$. Details of its numerical implementation can be found in [1]. In the table [1] on the obtained test set accuracy on the classification tasks and $R^2$ on the regression tasks. As we see from the table [1] after reducing the dimension of an input to $k = 2, 3$, we are still able to obtain good accuracy of prediction on a test set.

\(^2\)Since the role of the parameter $\sigma$ is similar to that of the bandwidth in the kernel density estimation, we use Silverman’s rule of thumb to set $\sigma = N^{-1/(n+4)}$. 

---

**Figure 1:**
(a) Visualization of outputs of the PCA and MMD-PCA methods. MMD-PCA (green line) tends to select a subcollection of points that sharply aligns along the local direction (i.e. the tangent line), whereas the first principal component (red line) reflects the global shape of data.

(b) The dependence of $\|P_t - P\|_F$ on $t$ for different values of parameters $\delta$ and $\lambda$. Left plot: $\|P_t - P\|_F$: $\delta = 0.05$, $\lambda = 20.0$, case I, $\delta = 0.05$, $\lambda = 100.0$, case II. Right plot: $\|P^* - P\|_F$ as a function of $\ln \sigma$.
![Image of a page from a document with text]

Table 1: The cross-validated accuracies/R² of KNN on 2 or 3-dimensional input representations.

| Dataset       | Method | Dimension | 2 | 3         | 2 | 3         | 2 | 3         | 2 | 3         |
|---------------|--------|-----------|---|-----------|---|-----------|---|-----------|---|-----------|
| Heart (acc)   | PCA    | 79.80     | 79.46 | 82.49     | 81.82 | 86.33     | 88.77 | 81.48     | 83.50 |
| Breast (acc)  | SIR    | 93.46     | 93.65 | 97.30     | 96.73 | 93.13     | 95.95 | 97.88     | 97.69 |
| Ionosphere (acc) | KDR    | 80.29     | 86.57 | 89.14     | 89.43 | 83.43     | 86.29 | 88.29     | 90.57 |
| Diabetas (R²) |        | 25.34     | 28.72 | 43.47     | 43.61 | 41.82     | 44.30 | 43.07     | 44.48 |
| Boston (R²)   |        | 56.42     | 67.12 | 76.03     | 74.29 | 77.88     | 79.97 | 73.21     | 77.88 |
| Wine (R²)     |        | 93.91     | 94.12 | 98.30     | 96.02 | 97.10     | 96.93 |

and the AS for the SDR-ORF is competitive in comparison with other methods. Note that all listed datasets are of moderate size and our Python scripts managed to compute an effective subspace in 3-5 minutes on a PC with GTX Titan X (Pascal), Intel Core i7-7700K (4.20 GHz), 64 GB RAM.

Experiments with the shadow/black removal. We made experiments with Yale B dataset [43], which is a popular benchmark for testing robust versions of PCA. That dataset contains images of 28 human subjects under 9 poses and 64 illumination conditions. Test images used in the experiments are cropped and re-sized to 168x192 images, making the dimensionality of every image 32256. Thus, each human subject corresponds to a collection of 32256-dimensional vectors that lie on some low-dimensional subspace of $\mathbb{R}^{32256}$. We search for this subspace, assuming that its dimension is either 1 or 5, using PCA and the Algorithm 2 with the kernel $K(x, y) = (x \cdot y)e^{-\frac{||x-y||^2}{2\sigma^2}}$ (which we simply call Gauss). Our experiments showed that behaviour of PCA and Gauss are quite similar that its dimension is either 1 or 5, using PCA and the Algorithm 2 with the kernel $K(x, y) = (x \cdot y)e^{-\frac{||x-y||^2}{2\sigma^2}}$ (which we simply call Gauss). Our experiments showed that behaviour of PCA and Gauss are quite similar.

Experiments with the background modeling. For these experiments we use the dataset for testing background estimation algorithms SBMnet [44]. The dataset contains frames of short videos and the frame of a background for each video (so called the ground truth). Spatial resolutions of the videos vary from 240x240 to 800x600. Thus, a collection of frames of every video is a set of high-dimensional vectors (with a dimension up to 480000) that, again, lie on a low dimensional subspace $\mathcal{L}$. We assume that the dimension of $\mathcal{L}$ is 5. We recover $\mathcal{L}$ using PCA and the Algorithm 2 for kernels $K(x, y) = \sum_{i=1}^{4}(\frac{x_i y_i}{n})^4$, $K(x, y) = (x \cdot y)e^{-\frac{||x-y||^2}{2\sigma^2}}$, $K(x, y) = (x \cdot y)e^{-\frac{||x-y||^2}{2\sigma^2}}$ and $K(x, y) = (x \cdot y)(1 + \frac{||x-y||^2}{n})^{-\frac{1}{2}}$ (which we simply call Kurtosis, Gauss, Laplace and Poisson respectively). Recall that, according to corollary 1 this algorithm is 2-approximating for Kurtosis. Subsequently, we compute the median of the vectors, projected onto $\mathcal{L}$, and define the latter to be the recovered background image (see Figure 4). Measures of consistency with the ground truth backgrounds are then calculated using Python scripts downloaded from [45]. Six measures are used: AGE (average of the gray-level absolute difference between a ground truth image and a computed background image), pEPs (percentage of pixels in a computed background whose value differs from the value of the corresponding pixel in a ground truth by more than a threshold), pCEPS (percentage of pixels whose 4-connected neighbors are also error pixels), MSSSIM (estimate of the perceived visual distortion), PSNR (Peak-Signal-to-Noise-Ratio, or $10 \log_{10} \left( \frac{L-1}{MSE} \right)$) where $L$ is the maximum number of grey levels and MSE is the mean squared error between a ground truth and a computed background images), CQM (Color image Quality Measure). Codes that compute listed metrics can be found in [45]. As shown on Table 3 experiments again demonstrated very similar behavior of PCA, Kurtosis, Gauss, Laplace and Poisson with very close accuracies of the background reconstruction. Best measures of consistency with the ground truth images were achieved for Gauss ($\alpha = 5.0, 25.0$) and Laplace ($\alpha = 5.0$). For a comparison with other methods based on low-rank approximation and an accuracy of a state-of-the-art method that was specifically tailored for that task [46]. On figure 5 one can see that background images computed by PCA and Gauss are almost identical, though Gauss is less than PCA to add local artefacts, such as blurs, noise etc.

The processing of the whole SBMnet dataset using PCA/Kurtosis/Gauss/ Laplace takes approximately the same time — 25 minutes on a cluster equipped with Intel Xeon Platinum 8168 Processors (33M Cache, 2.70 GHz) and 1TB RAM. The code is available on github to facilitate the reproducibility of our results.

Scalability of algorithms. A major practical limitation of the alternating scheme comes from the fact that it involves an optimization over a set of functions $f$, which in applications is either a feedforward neural network (as in our specifications of the AS for SDR-ORF, MMD-PCA, HM-MMMD-PCA) or a generative neural network (WGAN-PCA). A speed of optimization is also strongly dependant on the objective’s landscape. Thus, for large scale datasets, with a
Table 2: Original images (the first row) and their projections to 1-dimensional subspaces computed by PCA (the second row) and computed by Gauss (the third row).

Table 3: Original image, projected image and the difference between them (Gauss).

dimension of vectors $\gg 10^3$, and a sophisticated structure of a regression function (SDR-ORF) or a data distribution (MMD-PCA, HM-MMD-PCA, WD-PCA), the alternating scheme is substantially slower in comparison with other popular methods (such as PCA for the UDR, or SIR/KDR for the SDR).

In the special case of MMD-PCA (that includes HM-MMD-PCA), the approximate algorithm 2 can be used as a surrogate of the alternating scheme. It requires the same time as PCA and can be applied to datasets with a dimension of vectors $\sim 10^6 - 10^7$. Also, the Algorithm 2 can be used for an initialization of the alternating scheme.
Table 4: Original image, its projection, and their grayscale difference (Gauss).

9 Conclusions

We develop a new optimization framework for LDR problems. The alternating scheme for the optimization task demonstrates both the computational efficiency and the applicability to real-world data. The algorithm performs quite stably when we vary most of the hyperparameters, though it crucially depends on two parameters, the bandwidth of the “smoothing” kernel $M$, $\sigma$, and the penalty parameter $\lambda$. We believe that the MMD-PCA/HM-MMD-PCA/WD-PCA methods for UDR could be used as an alternative to PCA in study fields in which data demonstrate “heavy-tailed” and “non-Gaussian” behavior, such as financial applications or computer vision. Also, our formulation of SDR-ORF is free from any assumptions on the distribution of input-output pairs, which makes it an alternative to other methods of efficient subspace estimation. A more detailed report on these topics is a subject of future research.

References

[1] William B. Johnson, Joram Lindenstrauss, and Gideon Schechtman. Extensions of lipschitz maps into banach spaces. *Israel Journal of Mathematics*, 54(2):129–138, Jun 1986.

[2] John P. Cunningham and Zoubin Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, 16(89):2859–2900, 2015.

[3] P.-A. Absil, R. Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

[4] Laurent Schwartz. Théorie des distributions et transformation de fourier. *Annales de l’université de Grenoble*, 23:7–24, 1947.

[5] S. Soboleff. Méthode nouvelle à resoudre le problème de Cauchy pour les équations linéaires hyperboliques normales. *Rec. Math. Moscou, n. Ser.*, 1:39–71, 1936.

[6] Qiong Wang, Junbin Gao, and Hong Li. Grassmannian manifold optimization assisted sparse spectral clustering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3145–3153, 2017.

[7] Jiayao Zhang, Guangxu Zhu, Robert W. Heath, and Kaibin Huang. Grassmannian learning: Embedding geometry awareness in shallow and deep learning, 2018.

[8] Ziyu Wang, Frank Hutter, Masrour Zoghi, David Matheson, and Nando De Freitas. Bayesian optimization in a billion dimensions via random embeddings. *J. Artif. Int. Res.*, 55(1):361–387, January 2016.

[9] Paul G. Constantine. *Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies*. Society for Industrial and Applied Mathematics, USA, 2015.

[10] Massimo Fornasier, Karin Schnass, and Jan Vybiral. Learning functions of few arbitrary linear parameters in high dimensions. *Found. Comput. Math.*, 12(2):229–262, April 2012.

[11] Hemant Tyagi and Volkan Cevher. Learning non-parametric basis independent models from point queries via low-rank methods. *Applied and Computational Harmonic Analysis*, 37(3):389–412, 2014.
Input images

A background computed by PCA

A background computed by Gauss

Table 5: Computed backgrounds are almost identical, though noise is more often in PCA’s output.

[12] Meihong Wang, Fei Sha, and Michael I. Jordan. Unsupervised kernel dimension reduction. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, Advances in Neural Information Processing Systems 23, pages 2379–2387. Curran Associates, Inc., 2010.

[13] Ker-Chau Li. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 86(414):316–327, 1991.

[14] Ker-Chau Li. On principal hessian directions for data visualization and dimension reduction: Another application of stein’s lemma. Journal of the American Statistical Association, 87(420):1025–1039, 1992.

[15] R. Dennis Cook. Save: a method for dimension reduction and graphics in regression. Communications in Statistics - Theory and Methods, 29(9-10):2109–2121, 2000.

[16] R. Dennis Cook and Liliana Forzani. Principal fitted components for dimension reduction in regression. Statistical Science, 23(4):485–501, 2008.

[17] R. Dennis Cook and Liliana Forzani. Likelihood-based sufficient dimension reduction. Journal of the American Statistical Association, 104(485):197–208, 2009.
| Method       | AGE   | pEPs  | pCPSs | MSSSIM | PSNR  | CQM   |
|--------------|-------|-------|-------|--------|-------|-------|
| MSCL (SOTA)  | 5.9547| 0.0524| 0.0171| 0.9410 | 20.8952| 31.7049|
| BRTF [37]    | 9.5833| 0.1140| 0.0876| 0.9621 | 28.4655| 29.3246|
| GoDec [38]   | 11.5934| 0.1584| 0.0974| 0.8854 | 24.9954| 25.9955|
| PCA          | 9.3774| 0.0904| 0.0522| 0.9027 | 26.1549| 27.0502|

Table 6: Measures of consistency with the ground truth background image for the SBMnet dataset.

[18] Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *J. Mach. Learn. Res.*, 5:73–99, December 2004.

[19] Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Sufficient dimension reduction for high-dimensional regression and low-dimensional embedding: Tutorial and survey, 2021.

[20] A. Gretton, C. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, March 2012.

[21] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He. Fast and accurate matrix completion via truncated nuclear norm regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2117–2130, 2013.

[22] T. Oh, Y. Tai, J. Bazin, H. Kim, and I. S. Kweon. Partial sum minimization of singular values in robust pca: Algorithm and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):744–758, 2016.

[23] Q. Liu, Z. Lai, Z. Zhou, F. Kuang, and Z. Jin. A truncated nuclear norm regularization method based on weighted residual error for matrix completion. *IEEE Transactions on Image Processing*, 25(1):316–330, 2016.

[24] Bin Hong, Long Wei, Yao Hu, Deng Cai, and Xiaofei He. Online robust principal component analysis via truncated nuclear norm regularization. *Neurocomputing*, 175:216 – 222, 2016.

[25] Fumio Hiai. Concavity of certain matrix trace and norm functions. *Linear Algebra and its Applications*, 439(5):1568 – 1589, 2013.

[26] Yu Zhu and Peng Zeng. Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*, 101(476):1638–1651, 2006.

[27] Daniel Kapla, Lukas Fertl, and Efstatia Bura. Fusing sufficient dimension reduction with neural networks. *Computational Statistics and Data Analysis*, 168:107390, 2022.

[28] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(80):2287–2322, 2010.

[29] Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research*, 16(104):3367–3402, 2015.

[30] F.G. Friedlander and M.S. Joshi. *Introduction to the Theory of Distributions*. Cambridge University Press, 1998.

[31] Distributions: topology and sequential compactness. [https://cmouhot.files.wordpress.com/2010/02/main.pdf](https://cmouhot.files.wordpress.com/2010/02/main.pdf) Accessed: 2022-01-30.

[32] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[33] L. Khachiyan. On the complexity of approximating extremal determinants in matrices. *J. Complex.*, 11(1):138–153, mar 1995.

[34] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

[35] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3), June 2011.
[36] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha. R1-pca: Rotational invariant 11-norm principal component analysis for robust subspace factorization. In Proceedings of the 23rd International Conference on Machine Learning, ICML ‘06, page 281–288, New York, NY, USA, 2006. Association for Computing Machinery.

[37] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, Advances in Neural Information Processing Systems 23, pages 2496–2504. Curran Associates, Inc., 2010.

[38] Amit Deshpande and Rameshwar Pratap. One-Pass Additive-Error Subset Selection for lp Subspace Approximation. In Mikołaj Bojańczyk, Emanuela Merelli, and David P. Woodruff, editors, 49th International Colloquium on Automata, Languages, and Programming (ICALP 2022), volume 229 of Leibniz International Proceedings in Informatics (LIPIcs), pages 51:1–51:14, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[39] Amit Deshpande, Madhur Tulsiani, and Nisheeth K. Vishnoi. Algorithms and hardness for subspace approximation. In Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SODA ‘11, page 482–496, USA, 2011. Society for Industrial and Applied Mathematics.

[40] S. J. Bernau. The square root of a positive self-adjoint operator. Journal of the Australian Mathematical Society, 8(1):17–36, 1968.

[41] T. Hsing and R. Eubank. Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley Series in Probability and Statistics. Wiley, 2015.

[42] Maximal subset with rank k. https://math.stackexchange.com/questions/294404/maximal-subset-with-rank-k. Accessed: 2022-09-29.

[43] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(6):643–660, 2001.

[44] Pierre-Marc Jodoin, Lucia Maddalena, Alfredo Petrosino, and Yi Wang. Extensive benchmark and survey of modeling methods for scene background initialization. IEEE Transactions on Image Processing, 26(11):5244–5256, 2017.

[45] A dataset for testing background estimation algorithms. http://pione.dinf.usherbrooke.ca/ Accessed: 2022-08-22.

[46] Sajid Javed, Arif Mahmood, Thierry Bouwmans, and Soon Ki Jung. Background–foreground modeling based on spatiotemporal sparse subspace clustering. IEEE Transactions on Image Processing, 26(12):5840–5854, 2017.

[47] Qibin Zhao, Guoxu Zhou, Liqing Zhang, Andrzej Cichocki, and Shun-Ichi Amari. Bayesian robust tensor factorization for incomplete multivariate data. IEEE Transactions on Neural Networks and Learning Systems, 27(4):736–748, 2016.

[48] Tianyi Zhou and Dacheng Tao. Godec: Randomized low-rank sparse matrix decomposition in noisy case. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML’11, page 33–40, Madison, WI, USA, 2011. Omnipress.

[49] Advanced real analysis. https://warwick.ac.uk/fac/sci/masdoc/people/masdoc_alumni/davidmccormick/ara-v0.1.pdf. Accessed: 2022-01-30.

[50] S. Bochner. Vorlesungen über Fouriersche Integrale: von S. Bochner. Mathematik und ihre Anwendungen in Monographien und Lehrbüchern. Akad. Verl.-Ges., 1932.

[51] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information Theory, 39(3):930–945, May 1993.

[52] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[53] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 5767–5777. Curran Associates, Inc., 2017.

[54] Xiang Wei, Zixia Liu, Liqiang Wang, and Boqing Gong. Improving the improved training of wasserstein GANs. In International Conference on Learning Representations, 2018.
If we prove the second part of theorem we need the following lemma.

A.2 Proof of Theorem

Proof. The inclusion $\mathcal{G}_k' \subseteq \mathcal{G}_k^\star$ follows from a well-known fact that $\mathcal{S}(\mathbb{R}^k)$ is dense in $\mathcal{S}'(\mathbb{R}^k)$. I.e. for any $f \in \mathcal{S}'(\mathbb{R}^k)$ one can always find a sequence $\{f_i\} \subseteq \mathcal{S}(\mathbb{R}^k)$ such that $T_{f_i} \to f$. Therefore, for any $(f \otimes \delta^{n-k})_U \in \mathcal{G}_k'$ there is a sequence $\{(T_{f_i} \otimes \delta^{n-k})_U\} \subseteq \mathcal{G}_k$ such that $(T_{f_i} \otimes \delta^{n-k})_U \to (f \otimes \delta^{n-k})_U$. Thus, $\mathcal{G}_k' \subseteq \mathcal{G}_k^\star$.

Since $\mathcal{G}_k \subseteq \mathcal{G}_k'$, to prove $\mathcal{G}_k' = \mathcal{G}_k^\star$ it is enough to show that $\mathcal{G}_k'$ is sequentially closed.

We need a simple fact from a theory of distributions.

Lemma 2. If $T_i \to T$ and $\phi_i \to \phi$, then $(T_i, \phi_i) \to (T, \phi)$.

Proof of Lemma. Schwartz space $\mathcal{S}(\mathbb{R}^n)$ is a Fréchet space, therefore the Banach-Steinhaus theorem applies to $\mathcal{S}'(\mathbb{R}^n)$. Since $T_i \to T$, we have $\sup_{\phi} |\langle T_i, \phi \rangle| < \infty$ for any $\phi \in \mathcal{S}(\mathbb{R}^n)$. From the Banach-Steinhaus theorem, applied to a set $\{T_i\}_{i=1}^\infty$, we obtain for any $\epsilon > 0$, there is a neigbourhood $U$ of $0 \in \mathcal{S}(\mathbb{R}^n)$ such that $|\langle T_i, \phi \rangle| < \epsilon$ whenever $\phi \in U$. Thus, $|\langle T_i, \phi_i - \phi \rangle| < \epsilon$ for a large enough $i$. From that we conclude that $(T_i, \phi_i) \to (T, \phi)$.

For any $T \in \mathcal{S}'(\mathbb{R}^n)$ and $\psi \in \mathcal{S}(\mathbb{R}^{n-k})$, let us define $T^\psi \in \mathcal{S}'(\mathbb{R}^{n-k})$ as $\langle T^\psi, \phi \rangle = \langle T, \psi \otimes \phi \rangle$.

Suppose that $\{f_i\}_{i=1}^\infty \subseteq \mathcal{S}'(\mathbb{R}^k)$, $\{U_i\}_{i=1}^\infty$ are such that $(f_i \otimes \delta^{n-k})_{U_i} \to f$. We need to prove that $f \in \mathcal{G}_k'$. Since a set of orthogonal matrices is compact, then one can always find a subsequence $\{U_{n_i}\}$ such that $U_{n_i} \to \bar{U}$. Since $(f_{n_i} \otimes \delta^{n-k})_{U_{n_i}} \to f$ and $\phi(U_{n_i}, x) \to \phi(U, x)$ (for any fixed $\phi \in \mathcal{S}(\mathbb{R}^n)$), using Lemma 2 we obtain:

$$\langle f_{n_i} \otimes \delta^{n-k}, \phi \rangle = \langle (f_{n_i} \otimes \delta^{n-k})_{U_{n_i}}, \phi(U_{n_i}, x) \rangle \to \langle f, \phi(U, x) \rangle = \langle f^\psi_{U_T}, \phi(x) \rangle \quad (54)$$

Thus, we have $f_{n_i} \otimes \delta^{n-k} \to f_{U_T}$. From the last we see that $f_{n_i} \to f^\psi_{U_T}$ where $\psi$ is such that $\psi(0) = 1$. Therefore, $f_{U_T} = f^\psi_{U_T} \otimes \delta^{n-k}$ and $f = (f^\psi_{U_T} \otimes \delta^{n-k})_{U} \in \mathcal{G}_k'$.

A.2 Proof of Theorem

Proof of Theorem $\Rightarrow$. Let us prove that from $T = (f \otimes \delta^{n-k})_{U}, f \in \mathcal{S}'(\mathbb{R}^k), U^T U = I_n$ it follows that $\dim \text{span}_{\mathbb{R}}\{x_1 T, x_2 T, \ldots, x_n T\} \leq k$.

It is easy to see that $x_i[f \otimes \delta^{n-k}] = 0$ if $i > k$. If $U = [u_1, \ldots, u_n]^T$, then for $i > k$ we have $0 = (x_i[f \otimes \delta^{n-k}])_U = u_i^T x(f \otimes \delta^{n-k})_U = u_i^T xT$.

Thus, we have $n - k$ orthogonal vectors, $u_{k+1}, \ldots, u_n$, such that

$$[x_1 T \quad \cdots \quad x_n T] u_i = 0. \quad (55)$$

Using standard linear algebra we obtain there are at most $k'$ distributions $x_{i_1} T, \ldots, x_{i_{k'}} T, k' \leq k$ that form a basis of $\text{span}_{\mathbb{R}}\{x_i T\}_{i=1}^n$.

To prove the second part of theorem we need the following lemma.

Lemma 3. If $T \in \mathcal{S}'(\mathbb{R}^n)$ is such that $y_i T = 0$ for any $i > k$, then $T \in \mathcal{G}_k'$.

Proof of lemma. Recall from functional analysis, for $f \in \mathcal{S}'(\mathbb{R}^n)$, the tempered distribution $\frac{\partial f}{\partial x_i}$ is defined by the condition $\langle \frac{\partial f}{\partial x_i}, \phi \rangle = -\langle f, \frac{\partial \phi}{\partial x_i} \rangle$. Once the Fourier transform is applied, our lemma’s dual version is equivalent to the following formulation: if $\frac{\partial f}{\partial x_i} = 0, i > k$, then $f \in \mathcal{F}_{\mathbb{R}^n}$. Let us prove it in this formulation.

Recall that a set of infinitely differentiable functions with a compact support is denoted by $C_c^\infty(\mathbb{R})$. Suppose $\phi \in \mathcal{S}(\mathbb{R}^n)$ and $p \in C_c^\infty(\mathbb{R})$ are chosen in such a way that $\int_{-\infty}^{\infty} p(y_i)dy_i = 1$, $\text{supp} p \subseteq [A, B]$. Let us define:

$$r(x) = \int_{-\infty}^{x_i} \phi(x_i - y_i)dy_i - \int_{-\infty}^{x_i} p(y_i)dy_i \int_{-\infty}^{\infty} \phi(x_i - y_i)dy_i \quad (56)$$
Thus, $x_i > k$ implies:

$$\lim_{x_i \to \infty} x_i \frac{\partial^\beta \phi(x, y_i)}{\partial x_i^\beta} = \lim_{x_i \to \infty} x_i \frac{\partial^\beta \phi(x, y_i)}{\partial x_i^\beta} = \infty$$

The latter is also bounded, since the terms $x_i \frac{\partial^\beta \phi(x, y_i)}{\partial x_i^\beta}$ are bounded by the definition of $S(\mathbb{R}^n)$, $C^\infty_c(\mathbb{R})$. The boundedness of $\int_{-\infty}^{\infty} x_i \frac{\partial^\beta \phi(x, y_i)}{\partial x_i^\beta} dy_i$ is a consequence of the inequality $|x_i \frac{\partial^\beta \phi(x, y_i)}{\partial x_i^\beta}| \leq \frac{C}{1+y_i}$ (which holds because $\phi \in S(\mathbb{R}^n)$).

Analogously (when not a single derivative over $x_i$ is present):

$$x_i \frac{\partial^\beta}{\partial x_i^\beta} = x_i \int_{-\infty}^{x_i} \frac{\partial^\beta \phi(x, y_i)}{\partial x_i^\beta} dy_i - x_i \int_{x_i}^{x_i} \frac{\partial^\beta \phi(x, y_i)}{\partial x_i^\beta} dy_i = x_i (1 - \int_{-\infty}^{x_i} p(y_i)dy_i) \int_{-\infty}^{x_i} \frac{\partial^\beta \phi(x, y_i)}{\partial x_i^\beta} dy_i - x_i \int_{x_i}^{x_i} \frac{\partial^\beta \phi(x, y_i)}{\partial x_i^\beta} dy_i$$

The second term is 0 when $x_i \leq A$. It is also bounded when $x_i > A$ because $|x_i \frac{\partial^\beta \phi(x, y_i)}{\partial x_i^\beta}| \leq \frac{C}{(1+y_i)^{\alpha+1}}$.

Therefore,

$$\left| x_i \frac{\partial^\beta}{\partial x_i^\beta} \right| \leq \left| x_i \right| \frac{C}{(1+y_i)^{\alpha+1}}$$

The latter is bounded because $\lim_{x_i \to \infty} \frac{C}{(1+y_i)^{\alpha+1}} dy_i = 0$.

The first term is 0 when $x_i \geq B$ and for $x_i < B$ it satisfies:

$$\left| x_i \frac{\partial^\beta}{\partial x_i^\beta} \right| \leq \left| x_i \right| \frac{C}{(1+y_i)^{\alpha+1}}$$

The latter is also bounded, since $\lim_{x_i \to -\infty} \frac{C}{(1+y_i)^{\alpha+1}} dy_i = 0$.

Thus, $x_i \frac{\partial^\beta}{\partial x_i^\beta} \phi(x)$ is bounded and $r \in S(\mathbb{R}^n)$. Therefore, $\frac{\partial f}{\partial x_i} = 0$ implies:

$$\langle f, \frac{\partial r}{\partial x_i} \rangle = 0 \Rightarrow f[\phi] = f[p(x_i) \int_{-\infty}^{\infty} \phi(x, y_i) dy_i].$$

Since the sequence of arguments works for any $i > k$, we can apply them sequentially to initial $\phi \in S(\mathbb{R}^n)$ w.r.t. $x_{k+1}, \ldots, x_n$. Thus, for any $p_{k+1}, \ldots, p_n \in C_c(\mathbb{R})$ such that $\int_{-\infty}^{\infty} p_i(y_i)dy_i = 1$ we obtain:

$$f[\phi] = f[p_{k+1}(x_{k+1}) \cdots p_n(x_n) \int_{-\infty}^{\infty} \phi(x_{1:k}, x_{k+1:n}) dx_{k+1:n}]$$

Moreover, since $C^\infty_c(\mathbb{R})$ is dense in $S(\mathbb{R})$, we can assume that $p_{k+1}, \ldots, p_n \in S(\mathbb{R})$. For the inverse Fourier transform $T = F^{-1}[f]$ the latter condition becomes equivalent to:

$$\langle T, \phi \rangle = \langle T, p_{k+1}(x_{k+1}) \cdots p_n(x_n) \phi(x_{1:k}, 0_{k+1:n}) \rangle$$

for any $p_{k+1}, \ldots, p_n \in S(\mathbb{R})$ such that $p_i'(0) = 1$. Let us define $p_i'(x_i) = e^{-x_i^2}$. It is easy to check that $T = g \otimes \delta_n$ where $g \in S'(\mathbb{R})$, $\langle g, \psi \rangle = \langle T, e^{-|x|^2} \psi(x_{1:k}) \rangle$ for $\psi \in S(\mathbb{R}^k)$. Thus, $T \in G_k^c$ and lemma is proved.

**Proof of Theorem**: Let $\dim_{\mathbb{R}}(x_1 T, x_2 T, \cdots, x_n T) \leq k$, then

$$\dim_{\mathbb{R}}(v \in \mathbb{R}^n | x_1 T, \cdots, x_n T v = 0) \geq n - k.$$
Thus, there exist at least \( n - k \) orthonormal vectors \( \mathbf{v}_{k+1}, \ldots, \mathbf{v}_n \), such that \( [x_1 T, \ldots, x_n T] \mathbf{v}_i = 0 \). Therefore, \( [x_1 T, \ldots, x_n T] \mathbf{v}_i = (\mathbf{v}_i^T \mathbf{x}) T = 0 \).

Let us complete \( \mathbf{v}_{k+1}, \ldots, \mathbf{v}_n \) to form an orthonormal basis of \( \mathbb{R}^n \): \( \mathbf{v}_1, \ldots, \mathbf{v}_n \). Let us define a matrix \( V = [\mathbf{v}_1, \ldots, \mathbf{v}_n] \). It is easy to see that:

\[
(\mathbf{v}_i^T \mathbf{x}) V = (\mathbf{v}_i^T V \mathbf{x}) T = x_i T
\]

(65)

Since for \( i > k \) we have \((\mathbf{v}_i^T \mathbf{x}) T = 0\), then \( x_i T V = 0 \). Using lemma 3 we obtain \( T V \in G'_k \). Therefore, \((T V) V T = T \in G'_k \). Theorem proved.

B Structure of WD-PCA

Recall that \((\mathbb{R}^n, \| \cdot \|)\) is a Banach space and \( p \geq 1 \). Now, let us consider an optimization problem: for a given \( X \in \mathbb{R}^{n \times N} \) solve

\[
\|X - L\|_p \to \min_{\text{rank}(L) \leq k}
\]

(66)

where \( \| \cdot \|_p \) is a norm on \( \mathbb{R}^{n \times N} \) defined by \( \| [s_1, \ldots, s_N] \|_p \overset{\text{def}}{=} (\sum_{i=1}^N \| s_i \|_p^p)^{1/p} \).

The following simple theorem shows that the two tasks are connected so that the solution of one directly leads to another’s solution.

**Theorem 12.** Given data points \( \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \), let \( X = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{n \times N} \). Then,

\[
\min_{\nu \in P_k} W_p(\mu_{\text{data}}, \nu) = \frac{1}{N^p} \min_{Y \in \mathbb{R}^{n \times N}, \text{rank}(Y) \leq k} \|X - Y\|_p.
\]

(67)

Moreover, \( \min_{\nu \in P_k} W_p(\mu_{\text{data}}, \nu) \) is attained on \( \nu^* \), where \( \nu^* \) is a uniform distribution over \( \{Y_i\}_{i=1}^N \) and \( \arg \min_{Y \in \mathbb{R}^{n \times N}, \text{rank}(Y) \leq k} \|X - Y\|_p \) is attained on \( Y^* \), where \( Y^* = [y_1, \ldots, y_N] \in \arg \min_{Y \in \mathbb{R}^{n \times N}, \text{rank}(Y) \leq k} \|X - Y\|_p \).

**Proof.** Let us prove first that \( \inf_{\mu \in P_k} W_p(\mu_{\text{data}}, \mu) \leq \frac{1}{N^p} \|X - Y^*\|_p \) where

\[
Y^* = [y_1, \ldots, y_N] \in \arg \min_{Y \in \mathbb{R}^{n \times N}, \text{rank}(Y) \leq k} \|X - Y\|_p.
\]

(68)

Let \( \pi \) be a uniform distribution over \( \{(\mathbf{x}_i, y_i)\}_{i=1}^N \) and \( \mu^* \) be a uniform distribution over \( \{y_i\}_{i=1}^N \). Since \( \pi \in \Pi(\mu_{\text{data}}, \mu^*) \), we obtain \( W_p(\mu_{\text{data}}, \mu^*) \leq \frac{1}{N^p} \sum_{i=1}^N \|x_i - y_i\|^p_1 \leq \frac{1}{N^p} \|X - Y^*\|_p \). The support of \( \mu^* \) is k-dimensional, because \( \text{rank}(Y^*) \leq k \). Thus, we have \( \mu^* \in P_k \) and \( \inf_{\mu \in P_k} W_p(\mu_{\text{data}}, \mu) \leq W_p(\mu_{\text{data}}, \mu^*) \leq \frac{1}{N^p} \|X - Y^*\|_p \).

Now, if we prove the inverse inequality, i.e. \( \inf_{\mu \in P_k} W_p(\mu_{\text{data}}, \mu) \geq \frac{1}{N^p} \|X - Y^*\|_p \), this will imply that \inf_{\mu \in P_k} W_p(\mu_{\text{data}}, \mu) = \frac{1}{N^p} \|X - Y^*\|_p \) and therefore, \( \inf_{\mu \in P_k} W_p(\mu_{\text{data}}, \mu) = W_p(\mu_{\text{data}}, \mu^*) \). This will in the end give us \( \|X - Y\|_p \) is attained on \( \mu^* \).

Let \( \{\mu_t\}_{t=1}^\infty \) be such that \( \mu_t \in P_k \) and \( W_p(\mu_{\text{data}}, \mu_t) \to W_p(\mu_{\text{data}}, \mu) \to 0 \). Let \( L_t \) denote a k-dimensional support of \( \mu_t \) and \( P_t \) is a projection operator onto \( L_t \).

Let \( \mu^*_t \) be a uniform distribution over \( \{P_t \mathbf{x}_1, \ldots, P_t \mathbf{x}_N\} \), i.e. \( \mu^*_t(A) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{[P_t \mathbf{x}_i \in A]} \). It is easy to see that \( W_p(\mu^*_t, \mu_{\text{data}}) \leq W_p(\mu_t, \mu_{\text{data}}) \), because \( \mu^*_t \) and \( \mu_t \) share the same k-dimensional support \( L_t \), but the “transportation of mass” concentrated in point \( \mathbf{x}_i \) of the empirical distribution \( \mu_{\text{emp}} \) can be most optimally done by just moving it to \( P_t \mathbf{x}_i \) (i.e. to the closest point on \( L_t \)). Thus, we have \( \inf_{\mu \in P_k} W_p(\mu_{\text{data}}, \mu) \leq W_p(\mu_{\text{data}}, \mu^*_t) \leq W_p(\mu_{\text{data}}, \mu_t) \), and therefore, \( W_p(\mu_{\text{data}}, \mu^*_t) \to 0 \).

Since a set of projection operators is compact, one can always extract a subsequence \( \{P_{t_s}\}_{s=1}^\infty \) such that \( P_{t_s} \to P \). It is easy to see that \( \mu^*_t \to \mu^* \) (i.e. \( W_p(\mu^*_t, \mu^*) \to 0 \)) where \( \mu^* \) is a uniform distribution over \( \{P \mathbf{x}_1, \ldots, P \mathbf{x}_N\} \).

For that distribution we have

\[
W_p(\mu_{\text{data}}, \mu^*) = \lim_{s \to \infty} W_p(\mu_{\text{data}}, \mu^*_{t_s}) = \inf_{\mu \in P_k} W_p(\mu_{\text{data}}, \mu).
\]

(69)

Thus, the infimum is attained on \( \mu^* \).

It is easy to see that \( W_p(\mu_{\text{data}}, \mu^*) \to \frac{1}{N^p} \|X - PX\|_p \). Since \( \text{rank}(PX) \leq k \) we obtain \( W_p(\mu_{\text{data}}, \mu^*) \geq \frac{1}{N^p} \min_{Y \in \mathbb{R}^{n \times N}, \text{rank}(Y) \leq k} \|X - Y\|_p = \|X - Y^*\|_p \). This completes the proof.

\[\square\]
Note that in the case of $l_1$ norm and $p = 1$, i.e., $\|x\| = \sum_i |x_i|$, the task corresponds to the well-studied robust PCA problem \cite{robust_pca}. If, instead of the $l_1$-norm, we use the $l_2$-norm and $p \geq 1$, this leads to another task:

$$\|X - L\|_{p, 2} \rightarrow \min_{\text{rank}(L) \leq k}$$

(70)

where $\|(s_1, \ldots, s_N)\|_{p, 2} = (\sum_{i=1}^{N} \|s_i\|_2^p)^{1/p}$. This task has many applications in mathematics and is known as the subspace approximation problem \cite{subspace_approximation}.

C Proper kernels and proof of Theorem

C.1 Proof of Theorem \cite{kernel_theorem}

Let us first show that $\langle f | M | g \rangle$ is defined for any $f = (T_a \otimes \delta^{n-k})_U \in \mathcal{G}_k$ and $g = (T_b \otimes \delta^{n-k})_V \in \mathcal{G}_k$ where $a, b \in L_1(\mathbb{R}^k)$. We have

$$T_{f_e} = (T_a \otimes \delta^{n-k})_U \ast G^k_e = ((T_a \ast G^k_e) \otimes T_{G^{n-k}_e})_U$$

(71)

Let us denote $a_e = a \ast G^k_e$ and $b_e = b \ast G^k_e$. It is easy to see that

$$f_e = (a_e(x_{1:k})G^{n-k}_e(x_{k+1:n}))_U \in \mathcal{S}(\mathbb{R}^n).$$

(72)

From a well-known property of the Weierstrass transform we have

$$\|f_e\|_{L_1} = \|a_e\|_{L_1} \cdot \|G^{n-k}_e\|_{L_1} \leq \|a\|_{L_1}.$$  

(73)

From this we obtain that

$$|\langle f_e | M | g_e \rangle| \leq \max_{x,y} |M(x, y)| \|f_e\|_{L_1} \|g_e\|_{L_1} \leq \max_{x,y} |M(x, y)| \|a\|_{L_1} \|b\|_{L_1} < \infty.$$  

(74)

Thus, $\langle f_e | M | g_e \rangle$ is properly defined and

$$\langle f_e | M | g_e \rangle = \int_{\mathbb{R}^n \times \mathbb{R}^n} a^*_e(x_{1:k})G^{n-k}_e(x_{k+1:n})M(U^T x, V^T y)b_e(y_{1:k})G^{n-k}_e(y_{k+1:n})dxdy = \int_{\mathbb{R}^k \times \mathbb{R}^k} a^*_e(x_{1:k})M_e(x_{1:k}, y_{1:k})b_e(y_{1:k})dx_{1:k}dy_{1:k}$$

(75)

where

$$M_e(x_{1:k}, y_{1:k}) = \int_{\mathbb{R}^{n-k} \times \mathbb{R}^{n-k}} G^{n-k}_e(x_{k+1:n})M(U^T x, V^T y)G^{n-k}_e(y_{k+1:n})dx_{k+1:n}dy_{k+1:n}.$$

(76)

Let $U_k, V_k \in \mathbb{R}^{n \times n}$ be matrices that comprise the first $k$ rows of $U, V$ correspondingly and $n - k$ zero rows below. Also, let $L$ denote the Lipschitz constant for $M$ such that $|M(x, y) - M(x', y')| \leq L(|x - x'| + |y - y'|)$. For the function $M_e(x_{1:k}, y_{1:k})$ we have:

$$|M_e(x_{1:k}, y_{1:k}) - M(U^T_k x, V^T_k y)| \leq \left| \int_{\mathbb{R}^{2n-2k}} G^{n-k}_e(x_{k+1:n})(M(U^T x, V^T y) - M(U^T_k x, V^T_k y))dx_{k+1:n}dy_{k+1:n}\right| \leq L\left| \int_{\mathbb{R}^{2n-2k}} G^{n-k}_e(x_{k+1:n})\left(|(U - U_k)^T x| + |(V - V_k)^T y|\right)G^{n-k}_e(y_{k+1:n})dx_{k+1:n}dy_{k+1:n}\right| =$$

$$= L\left| \int_{\mathbb{R}^{2n-2k}} G^{n-k}_e(x_{k+1:n})\left(|x_{k+1:n}| + |y_{k+1:n}|\right)G^{n-k}_e(y_{k+1:n})dx_{k+1:n}dy_{k+1:n}\right| = \left(2L\int_{\mathbb{R}^{n-k}} |x_{k+1:n}|G^{n-k}_e(x_{k+1:n})dx_{k+1:n}\right) =$$

$$= 2L \int_{\mathbb{R}^{n-k}} |x_{k+1:n}|G^{n-k}_e(x_{k+1:n})dx_{k+1:n} \leq 2Le^{n-k} \int_{\mathbb{R}^{n-k}} |x_{k+1:n}|G^{n-k}_1(x_{k+1:n})dx_{k+1:n}.$$
Thus, there exists bounded $\tilde{M}(x_{1:k}, y_{1:k}) = M(U^T x, V^T y)$ such that

$$M_\sigma(x_{1:k}, y_{1:k}) \xrightarrow{\sigma \to 0} \tilde{M}(x_{1:k}, y_{1:k}) \text{ in } L_\infty(\mathbb{R}^{2k}).$$

(78)

Further we assume that $\epsilon > 0$ is small enough, so that $M_\epsilon(x_{1:k}, y_{1:k}) \leq C = 2\max |M(x, y)|$. Now we have:

$$|\langle f_\epsilon|M|g_\epsilon \rangle - \int_{\mathbb{R}^k \times \mathbb{R}^k} a^*(x_{1:k})\tilde{M}(x_{1:k}, y_{1:k})b(y_{1:k})dx_{1:k}dy_{1:k}| =$$

$$|\int_{\mathbb{R}^k \times \mathbb{R}^k} (a^*(x_{1:k})M_\epsilon(x_{1:k}, y_{1:k})b(y_{1:k}) - a^*(x_{1:k})\tilde{M}(x_{1:k}, y_{1:k})b(y_{1:k}))dx_{1:k}dy_{1:k} +$$

$$\int_{\mathbb{R}^k \times \mathbb{R}^k} M_\epsilon(x_{1:k}, y_{1:k})a^*(x_{1:k})\left(b(y_{1:k}) - b(y_{1:k})\right)dx_{1:k}dy_{1:k} +$$

$$\int_{\mathbb{R}^k \times \mathbb{R}^k} a^*(x_{1:k})b(y_{1:k})\left(M_\epsilon(x_{1:k}, y_{1:k}) - \tilde{M}(x_{1:k}, y_{1:k})\right)dx_{1:k}dy_{1:k}| \leq$$

$$C\|a_\epsilon\|_{L_1}\|b_\epsilon - b\|_{L_1} + C\|b\|_{L_1}\|a_\epsilon - a\|_{L_1} + \|a^*(x_{1:k})b(y_{1:k})\|_{L_1}\|M_\epsilon - \tilde{M}\|_{L_\infty}.$$

(79)

It is well-known (e.g. see Theorem 2.25 from [49]) that $\|a_\epsilon - a\|_{L_p}, \|b_\epsilon - b\|_{L_p} \to 0$, $\|a_\epsilon\|_{L_1} \leq \|a\|_{L_1}$ and $\|M_\epsilon - \tilde{M}\|_{L_\infty} \to 0$. Thus, $\lim_{\epsilon \to 0}\langle f_\epsilon|M|g_\epsilon \rangle$ exists and $\langle f|M|g \rangle$ is defined.

Let us now prove that rank $M_f \leq k$. The function $f \in G_k$ is such that $f = (T_g \otimes \delta^{n-k})_U$ where $\{x, g\}_{i=1}^k \subseteq L_1(\mathbb{R}^k)$ and $U = [w_1, \ldots, w_n]$ is an orthogonal matrix. By construction,

$$\langle x, f \rangle|_{M|x,j} = \langle (x, f)_U^T | [M(U^T x, U^T y)](x, f)^U \rangle = \langle (w_i^T x T_g \otimes \delta^{n-k})|M(U^T x, U^T y)|w_j^T x T_g \otimes \delta^{n-k} \rangle. \quad (80)$$

Let us now denote $V = [u_1, \ldots, u_n] \in \mathbb{R}^{k \times n}$ a submatrix of $U$ in which only first $k$ rows of $U$ are present. Then, the latter integral is equal to

$$\int_{\mathbb{R}^k \times \mathbb{R}^k} u_i^T x_{1:k} y_{1:k} u_j g(x_{1:k})^* M(V^T x_{1:k}, V^T y_{1:k}) g(y_{1:k}) dx_{1:k} dy_{1:k} = u_i^T B u_j \quad (81)$$

where

$$B = \{\langle x, g(x_{1:k}) \rangle \}_{i=1}^k \in L_1(\mathbb{R}^k).$$

(82)

is the Gram matrix of the collection $\{x, g(x_{1:k})\}_{i=1}^k \subseteq L_1(\mathbb{R}^k)$.

Obviously, rank $M_f = \text{rank } \left[ \text{Re } u_i^T B u_j \right]_{i=1}^n \leq \text{rank } V^T (\text{Re } B) V \leq \text{rank } V = k.$

**D  Proofs of Theorem 5 and 6**

For any $f = (T_i \otimes \delta^{n-k})_U \in G_k$ and $\sigma > 0$, let us define $f_\sigma$ as

$$T_{f_\sigma} = (T_i \otimes \delta^{n-k})_U \ast G_\sigma = (T_{l_\sigma} \otimes T_{G_\sigma^{n-k}})_U$$

$$f_\sigma = (l_\sigma(x_{1:k})G_\sigma^{n-k}(x_{k+1:n}))_U$$

$$l_\sigma = l \ast G_\sigma^2. \quad (83)$$

We have $T_{f_\sigma} \xrightarrow{\sigma \to 0} (T_i \otimes \delta^{n-k})_U$ as $\sigma \to +0$.

**Lemma 4.** For any $f \in G_k$, $\lim_{\sigma \to +0}(x_j f_\sigma|M|x_j f_\sigma) = 0$, for any $(i, j) \notin \{1, \ldots, k\}^2$, and $\sup_{\sigma \in [0,1]}(x_j f_\sigma|M|x_j f_\sigma) < \infty$, for any $(i, j) \in \{1, \ldots, k\}^2$. 

26
Proof. W.l.o.g. we can assume that \( f = T_l \otimes \delta^{n-k}, l \in \mathcal{S}(\mathbb{R}^k) \). If \( i > k, j \leq k \) we have

\[
\langle x_i f_\sigma | M | x_j f_\sigma \rangle = \frac{1}{(2\pi\sigma^2)^{n-k}} \iint_{\mathbb{R}^n \times \mathbb{R}^n} x_i y_j e^{-\frac{|x_{k+1:n}|^2}{2\sigma^2}} l_\sigma(x_{1:k}) M(x, y) e^{-\frac{|y_{k+1:n}|^2}{2\sigma^2}} l_\sigma(y_{1:k}) \, dx \, dy = \int_{\mathbb{R}^n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{|x_{k+1:n}|^2}{2\sigma^2}} l_\sigma(x_{1:k}) P(x) \, dx
\]  

where \( P(x) = \int_{\mathbb{R}^n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{|y_{k+1:n}|^2}{2\sigma^2}} l_\sigma(y_{1:k}) \, dy \). Using the Hölder inequality we obtain

\[
\|\langle x_i f_\sigma | M | x_j f_\sigma \rangle\| \leq \frac{1}{\sqrt{2\pi\sigma^2}} \| x_i \|_{L_1(\mathbb{R}^n)} \| y_j \|_{L_1(\mathbb{R}^n)} \| l_\sigma \|_{L_1(\mathbb{R}^n)} \| P \|_{L_\infty(\mathbb{R}^n)} = \gamma \frac{1}{\sqrt{2\pi\sigma^2}} \| x_i \|_{L_1(\mathbb{R}^n)} \| y_j \|_{L_1(\mathbb{R}^n)} \| l_\sigma \|_{L_1(\mathbb{R}^n)} \| P \|_{L_\infty(\mathbb{R}^n)}
\]

Since \( |M(x, y)| \leq \gamma \) for some \( \gamma \), we have

\[
|P(x)| \leq \gamma \frac{1}{\sqrt{2\pi\sigma^2}} \| x_i \|_{L_1(\mathbb{R}^n)} \| y_j \|_{L_1(\mathbb{R}^n)} \| l_\sigma \|_{L_1(\mathbb{R}^n)} \| P \|_{L_\infty(\mathbb{R}^n)} = \gamma \frac{1}{\sqrt{2\pi\sigma^2}} \| x_i \|_{L_1(\mathbb{R}^n)} \| y_j \|_{L_1(\mathbb{R}^n)} \| l_\sigma \|_{L_1(\mathbb{R}^n)} \gamma \| y_j \|_{L_1(\mathbb{R}^n)} \| l_\sigma \|_{L_1(\mathbb{R}^n)}
\]

Thus,

\[
\|\langle x_i f_\sigma | M | x_j f_\sigma \rangle\| \leq \gamma \frac{1}{\sqrt{2\pi\sigma^2}} \| x_i \|_{L_1(\mathbb{R}^n)} \| y_j \|_{L_1(\mathbb{R}^n)} \| l_\sigma \|_{L_1(\mathbb{R}^n)} \gamma \| y_j \|_{L_1(\mathbb{R}^n)} \| l_\sigma \|_{L_1(\mathbb{R}^n)}. \tag{87}
\]

Using \( \| l_\sigma \|_{L_1(\mathbb{R}^k)} \to \| l \|_{L_1(\mathbb{R}^k)} \) \( \sigma \to 0 \), \( \| y_j \|_{L_1(\mathbb{R}^n)} \to \| y_j \|_{L_1(\mathbb{R}^n)} \) \( \sigma \to 0 \), we see the boundedness of \( \| l_\sigma \|_{L_1(\mathbb{R}^k)} \gamma \| y_j \|_{L_1(\mathbb{R}^n)} \| l_\sigma \|_{L_1(\mathbb{R}^n)} \) and proceed

\[
\leq C \frac{1}{\sqrt{2\pi\sigma^2}} \| x_i \|_{L_1(\mathbb{R}^n)} \| y_j \|_{L_1(\mathbb{R}^n)} \| l_\sigma \|_{L_1(\mathbb{R}^n)} \gamma \| y_j \|_{L_1(\mathbb{R}^n)} \| l_\sigma \|_{L_1(\mathbb{R}^n)}. \tag{88}
\]

It is easy to see that \( \frac{1}{\sqrt{2\pi\sigma^2}} \| x_i \|_{L_1(\mathbb{R}^n)} \to 0 \) as \( \sigma \to 0 \) and \( \langle x_i f_\sigma | M | x_j f_\sigma \rangle \to 0 \).

Similarly, we can prove that \( \langle x_i f_\sigma | M | x_j f_\sigma \rangle \to 0 \) if \( i, j > k \).

The entries of the main \( k \times k \) minor \( \langle (x_i f_\sigma | M | x_j f_\sigma) \rangle_{1 \leq i, j \leq k} \) are bounded, due to

\[
\text{Tr} M_{f_\sigma} = \frac{1}{(2\pi\sigma^2)^{n-k}} \iint_{\mathbb{R}^n \times \mathbb{R}^n} x \cdot y e^{-\frac{|x_{k+1:n}|^2}{2\sigma^2}} l_\sigma(x_{1:k}) M(x, y) e^{-\frac{|y_{k+1:n}|^2}{2\sigma^2}} l_\sigma(y_{1:k}) \, dx \, dy \leq \gamma \frac{1}{(2\pi\sigma^2)^{n-k}} \iint_{\mathbb{R}^n \times \mathbb{R}^n} (|x_{1:k} \cdot y_{1:k}| + |x_{k+1:n} \cdot y_{k+1:n}|) e^{-\frac{|x_{k+1:n}|^2}{2\sigma^2} + \frac{|y_{k+1:n}|^2}{2\sigma^2}} l_\sigma(x_{1:k}) l_\sigma(y_{1:k}) \, dx \, dy \leq \gamma \int_{\mathbb{R}^n \times \mathbb{R}^n} |x_{1:k} \cdot y_{1:k}| l_\sigma(x_{1:k}) l_\sigma(y_{1:k}) \, dx \, dy \, \| l_\sigma \|^2_{L_1(\mathbb{R}^{n-k})} \leq \gamma \sum_{j=1}^k \| y_j l_\sigma \|^2_{L_1(\mathbb{R}^{n-k})} + \| l_\sigma \|^2_{L_1(\mathbb{R}^{n-k})} (n-k) \sigma^2. \tag{89}
\]

Again, using \( \| l_\sigma \|_{L_1(\mathbb{R}^k)} \to \| l \|_{L_1(\mathbb{R}^k)} \) \( \sigma \to 0 \), \( \| y_j l_\sigma \|_{L_1(\mathbb{R}^n)} \to \| y_j l \|_{L_1(\mathbb{R}^n)} \) \( \sigma \to 0 \), we obtain the boundedness of RHS.

Corollary 2. For any \( f \in \mathcal{G}_k \), \( \lim_{\sigma \to 0} R(f_\sigma) = 0 \).

Proof. W.l.o.g. we can assume that \( f = T_l \otimes \delta^{n-k}, l \in \mathcal{S}(\mathbb{R}^k) \). By lemma, all entries of \( M_{f_\sigma} \) except those of the main \( k \times k \) minor approach 0 as \( \sigma \to 0 \). This means that \( \lim_{\sigma \to 0} Q(f_\sigma) = 0 \), where \( Q(f_\sigma) = \sum_{i=k+1}^n \langle x_i f_\sigma | M | x_i f_\sigma \rangle \).
Let $v_1, \ldots, v_n$ be unit eigenvectors of $M_{f_\sigma}$ corresponding to the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$, $P = \sum_{i=k+1}^n e_i e_i^T$, then

$$R(f_\sigma) = \sum_{i=k+1}^n \lambda_i = \min_{p_i \in [0,1], \sum_p_i=n-k} \sum_{i=1}^n \lambda_i p_i \leq \sum_{i=1}^n \lambda_i \text{Tr} (P v_i v_i^T P) = \text{Tr} (PM_{f_\sigma}P) = Q(f_\sigma) \tag{90}$$

Since $R(f_\sigma) \leq Q(f_\sigma)$, we obtain $\lim_{\sigma \to 0} R(f_\sigma) = 0$.

### D.0.1 Proof of Theorem 5

**Proof.** Suppose that a sequence $\{f_i\}_{i=1}^\infty \subseteq S(\mathbb{R}^n)$ regularly solves (7) and $T \in \text{Lim } f_i$. W.l.o.g. we can assume that $T_{f_i} \to^* T$ and $\text{Tr}(M_{f_i})$ is bounded and $I(f_i) + \lambda_i R(f_i) \leq \inf_{f \in S(\mathbb{R}^n)} I(f) + \lambda_i R(f) + \epsilon_i, \epsilon_i \to 0$. Below we use continuity of $I$ and corollary 2

$$\inf_{f \in S(\mathbb{R}^n)} I(f) + \lambda_i R(f) \leq \inf_{f \in G_k} I(f) + \lambda_i R(f) \leq \inf_{f \in G_k} \lim_{n \to \infty} I(f_i) + \lambda_i R(f_i) \leq \inf_{f \in G_k} I(f) \tag{91}$$

from which we conclude that $\lambda_i R(f_i) \leq \inf_{f \in G_k} I(f) + \epsilon_i$ and, therefore, $R(f_i) \rightarrow^\infty 0$.

For each $l$, let us define $P_l$ as the projection operator to a subspace spanned by first principal components of the matrix $\sqrt{M_{f_i}}$, i.e.

$$P_l = \sum_{i=1}^k v_i^l v_i^T \tag{92}$$

where $v_1^l, \ldots, v_k^l$ are orthonormal eigenvectors that correspond to $k$ largest eigenvalues of $\sqrt{M_{f_i}}$. From the Eckart-Young-Mirsky theorem we see that $R(f_i) = \|\sqrt{M_{f_i}} - P_l \sqrt{M_{f_i}}\|_F^2$. Since a set of all projection operators $\{P \in \mathbb{R}^{n \times n} | P^2 = P, P^T = P\}$ is a compact subset of $\mathbb{R}^{n \times n}$, one can always find a projection operator $P = \sum_{i=1}^k v_i v_i^T$ and a growing subsequence $\{f_i\}$ such that $\|P_{f_i} - P\|_F \to 0$ as $s \to \infty$. Thus, for the subsequence $\{f_i\}$ we have

$$\|\sqrt{M_{f_i}} - P_{f_i}\|_F = \|\sqrt{M_{f_i}} - P_{f_i} \sqrt{M_{f_i}} + P_{f_i} \sqrt{M_{f_i}} - P_{f_i}\|_F \leq \|\sqrt{M_{f_i}} - P_{f_i} \sqrt{M_{f_i}}\|_F + \|P_{f_i} - P\|_F \sqrt{\text{Tr}(M_{f_i})} \tag{93}$$

and using the boundedness of $\text{Tr}(M_{f_i})$ we obtain $\|\sqrt{M_{f_i}} - P_{f_i}\|_F \to 0$.

Since $\|\sqrt{M_{f_i}} - P_{f_i}\|_F \to 0$, let us complete $v_1, \ldots, v_k$ to an orthonormal basis and make the change of variables $y_i = v_i^T x$. Let us denote $V = [v_1, \ldots, v_n]$ and let $V^T = [w_1, \ldots, w_n]$. Then, after that change of variables any function $f(x)$ corresponds to $f'(y) = f(Vy)$ and the kernel $M$ corresponds to $M'(y, y') = M(Vy, Vy')$. After we apply that change of variables in the integral expression of $\langle x, f | M | x, f \rangle$, we obtain

$$\langle x, f | M | x, f \rangle = \langle x^T f | M' | x^T f \rangle = w_i^T \{y_j f'(M') y_j f' \}_{n \times n} w_j \Rightarrow \text{Re} \langle x, f | M | x, f \rangle = w_i^T \{y_j f'(M') y_j f' \}_{n \times n} w_j \tag{94}$$

I.e. $M_f = V M'_{f_i} V^T$, or $M'_{f_i} = V^T M_f V$. Note that $P = VT_n^k V^T$ where $T_n^k$ is a diagonal matrix whose main $k \times k$ minor is the identity matrix, and all other entries are zeros. Using the fact that the Frobenius norm of orthogonally similar matrices are equal and the identity $V^T \sqrt{M_{f_i}} V = \sqrt{V^T M_{f_i} V}$, we obtain

$$\|\sqrt{M_{f_i}} - P_{f_i}\|_F = \|V^T \sqrt{M_{f_i}} V - V^T P_{f_i} V\|_F = \|V^T M_{f_i} V - V^T P_{f_i} V\|_F = \|M_{f_i} - T_n^k \sqrt{M_{f_i}}\|_F \tag{95}$$

Thus, the property $\|\sqrt{M_{f_i}} - P_{f_i}\|_F \to 0$ implies

$$\text{Re} \langle y_i f'_i | M' | y_j f'_j \rangle \to 0, \text{ if } i > k. \tag{96}$$

Moreover, for $i = j$ we have $\text{Re} \langle y_i f'_i | M' | y_j f'_j \rangle = \langle y_i f'_i | M' | y_j f'_j \rangle$. It is easy to see that after the change of variables we still have $f'_i \to^* T V$. Since $f_i \in S(\mathbb{R}^n)$, we have $y_i f'_i \in S(\mathbb{R}^n)$ and, therefore, $y_i f'_i \in L_2(\mathbb{R}^n)$.  

28
Let us treat now \( M' \) as an operator \( O_{M'} : L_2(\mathbb{R}^n) \to L_2(\mathbb{R}^n), O_{M'}[f](x) = \int_{\mathbb{R}^n} M'(x, y) f(y) dy \). Let us take any function \( \phi \in L_2(\mathbb{R}^n) \) such that \( \psi = O_{M'}[\phi] \in S(\mathbb{R}^n) \). Since \( O_{M'} \) is a strictly positive self-adjoint operator, by the Cauchy-Schwarz inequality, we obtain
\[
|\langle y_i f'_i, O_{M'}[\phi] \rangle| \leq \sqrt{\langle y_i f'_i, M' y_i f'_i \rangle} \sqrt{\langle \phi, O_{M'}[\phi] \rangle}.
\] (97)
Therefore, for any \( \psi \in \text{Range} [O_{M'}] \cap S(\mathbb{R}^n) \) and \( i > k \) we have \( \lim_{s \to \infty} \langle y_i f'_i, \psi \rangle = \lim_{s \to \infty} \langle f'_i, y_i \psi \rangle = 0 \).

Since \( f'_i \to^* T \psi \) we obtain \( (T \psi, \psi) = \langle y_i T \psi, \psi \rangle = 0 \) for any \( \psi \in \text{Range} [O_{M'}] \cap S(\mathbb{R}^n) \). But the denseness of \( \text{Range} [O_{M'}] \cap S(\mathbb{R}^n) \) in \( S(\mathbb{R}^n) \) implies that \( y_i T \psi = 0 \).

Using lemma 4 and \( (T \psi, \psi) = T \psi \) we obtain \( T \in \mathcal{G}'_k \). Thus, we proved that \( T_{f_i} \to T \in \mathcal{G}'_k \).

Since \( I(f_i) \leq I(f) + \lambda_i R(f_i) \leq \inf_{f \in \mathcal{V}_k} I(f) + \epsilon_i \) and \( I \) is continuous, we finally get that \( I(T) \leq \inf_{f \in \mathcal{V}_k} I(f) \), i.e. \( T \in \text{Arg min}_{f \in \mathcal{V}_k} I(f) \).

\[\square\]

D.0.2 Proof of Theorem 10

Proof. Suppose \( f^* \in \text{Arg min}_{f \in \mathcal{V}_k} I(f) \cap \mathcal{B}_k \), i.e. \( f^* \in \mathcal{B}_k \) and \( I(f^*) = \min_{f \in \mathcal{V}_k} I(f) \). Since \( f^* \in \mathcal{B}_k \), there exists a sequence \( \{s^i\} \subseteq \mathcal{G}_k \) such that \( T_{s^i} \to^* f^* \) and \( \sup \text{Tr} M_{s^i} < \infty \).

Let us define \( s^{i \sigma}_i \in S(\mathbb{R}^n) \) as \( T_{s^{i \sigma}_i} = T_{s^i} \ast G^{n_\sigma} \). Since \( \lim_{\sigma \to 0} R(s^{i \sigma}_i) = 0 \) (lemma 5), there exists \( \sigma_i > 0 \), such that \( R(s^{i \sigma}_i) < \frac{1}{4} \) whenever \( 0 < \sigma \leq \sigma_i \). Also, by definition \( \text{Tr} M_{s^i} = \lim_{\sigma \to 0} \text{Tr} M_{s^{i \sigma}_i} \). Therefore, there exists \( \sigma'_i > 0 \), such that \( \text{Tr} M_{s^{i \sigma'_i}_i} < \text{Tr} M_{s^i} + 1 \) whenever \( 0 < \sigma \leq \sigma'_i \).

If we set \( \sigma_i^* = \min \{ \sigma_i, \sigma'_i, \frac{1}{7} \} \), then a sequence \( \{s^{i \sigma_i^*}_i\} \subseteq S(\mathbb{R}^n) \) satisfies
\[
\lim_{i \to \infty} R(s^{i \sigma_i^*}_i) = 0, \sup \text{Tr} M_{s^{i \sigma_i^*}_i} < \infty
\] (98)
and (using lemma 5) \( T_{s^{i \sigma_i^*}_i} \to^* f^* \).

Due to the continuity of \( I \) we have \( \lim_{i \to \infty} I(s^{i \sigma_i^*}_i) = I(f^*) \). Now we set \( f_i = s^{i \sigma_i^*}_i \), \( \lambda_i = \frac{1}{\sqrt{R(f_i)}} \) and we obtain the needed sequence:
\[
\lim_{i \to \infty} I(f_i) = \lim_{i \to \infty} I(f_i) + \lambda_i R(f_i) = I(f^*) \quad \text{lim} \quad \lambda_i = +\infty,
\] (99)
where \( \text{Tr} M_{f_i} \) is bounded. It remains to check that our sequence regularly solves (7), i.e. \( \lim_{i \to \infty} \inf_{f \in S(\mathbb{R}^n)} I(f) + \lambda_i R(f) = I(f^*) \) (this will imply \( \lim_{i \to \infty} I(f_i) + \lambda_i R(f_i) = \inf_{f \in S(\mathbb{R}^n)} I(f) = 0 \)). The inequality in one direction is obvious,
\[
\inf_{f \in S(\mathbb{R}^n)} I(f) + \lambda_i R(f) \leq \inf_{f \in \mathcal{V}_k} \lambda_i R(f) \leq \inf_{f \in \mathcal{V}_k} I(f) = I(f^*).
\] (100)

Let us prove the inverse inequality.

Since \( \text{rsol} (I(f), R(f)) \neq \emptyset \), there exists \( \{\tilde{f}_i\} \subseteq S(\mathbb{R}^n) \) such that
\[
I(\tilde{f}_i) + \lambda_i R(\tilde{f}_i) \leq \inf_{f \in S(\mathbb{R}^n)} I(f) + \lambda_i R(f) + \epsilon_i, \quad \lim_{i \to +\infty} \lambda_i = +\infty, \quad \lim_{i \to +\infty} \epsilon_i = 0, \text{Tr} M_{\tilde{f}_i} < \infty
\] (101)
and \( a = \lim_{i \to +\infty} T_{\tilde{f}_i} \). From theorem 5 we obtain \( a \in \text{Arg min}_{f \in \mathcal{V}_k} I(f) \).

One can always find a subset \( \{\lambda_{d_i}\} \subseteq \{\lambda_i\} \) such that \( \lambda_{d_i} < \lambda_i, \lambda_{d_i} \to \infty \) and obtain
\[
\inf_{f \in S(\mathbb{R}^n)} I(f) + \lambda_i R(f) \geq \inf_{f \in S(\mathbb{R}^n)} I(f) + \lambda_{d_i} R(f) \geq I(\tilde{f}_{d_i}) + \lambda_{d_i} R(\tilde{f}_{d_i}) - \epsilon_{d_i} \geq I(\tilde{f}_{d_i}) - \epsilon_{d_i}.
\] (102)

Therefore,
\[
\lim_{i \to \infty} \inf_{f \in S(\mathbb{R}^n)} I(f) + \lambda_i R(f) \geq \lim_{i \to \infty} I(\tilde{f}_{d_i}) - \epsilon_{d_i} = I(a) = \inf_{f \in \mathcal{V}_k} I(f) = I(f^*).
\] (103)

This proves that \( \{f_i\} \) regularly solves (7) and \( \lim_{i \to \infty} f_i = f^* \) i.e. \( f^* \in \text{rsol} (I(f), R(f)) \).
E  The alternating scheme in the dual space for $M(x, y) = \zeta(x - y)$

When $M(x, y) = \zeta(x - y)$, the alternating scheme allows for a reformulation in the dual space. By this we mean that in Scheme we substitute $\hat{\phi}_t$ for the original $\phi_t$. If the primal Scheme deals with operators $S_{\phi}, S_{\phi_{t-1}}$, the dual version deals with vectors of functions $\sqrt{\zeta \frac{\partial \hat{\phi}_t}{\partial x}}, \sqrt{\zeta \frac{\partial \hat{\phi}_{t-1}}{\partial x}}$. The substitution is based on the following simple fact:

**Theorem 13.** If $M(x, y) = \zeta(x - y), \zeta, \hat{\zeta} \in C(\mathbb{R}^n)$ and $\forall x \cdot \hat{\zeta}(x) > 0$, then there exist constants $c_1$ and $c_2$ such that $\|S_{\phi} - P_{t-1}S_{\phi_{t-1}}\|^2 = c_1 \| \frac{\partial \hat{\phi}_{t-1}}{\partial x} - P_{t-1} \frac{\partial \hat{\phi}_{t-1}}{\partial x} \|^2 \| \zeta_{<2,\hat{\zeta}}(\mathbb{R}^n) \|$ and $(x, j)M[x_j] = c_2 (\frac{\partial \hat{F}}{\partial x}, \frac{\partial \hat{F}}{\partial x})_{L_2,\hat{\zeta}}(\mathbb{R}^n)$

**Proof.** Let $f : \mathbb{R}^n \rightarrow \mathbb{C}$ be such that $\|x_i f\|_{L_2(\mathbb{R}^n)} < \infty$.

$$O_M[\psi] = \zeta * \psi \Rightarrow F\{O_M[\psi]\} \propto \zeta \hat{\psi} \Rightarrow F\{\sqrt{O_M[\psi]}\} \propto \sqrt{\zeta} \hat{\psi} \Rightarrow$$

$$S_f[\psi] = \text{Re} \langle x_i, \sqrt{O_M[\psi]} \rangle \propto \text{Re} \langle F\{x_i f\}, F\{\sqrt{O_M[\psi]}\} \rangle \propto \text{Re} \langle i\sqrt{\zeta} \frac{\partial f}{\partial x_i}, \sqrt{\zeta} \hat{\psi} \rangle = \text{Re} \langle i\sqrt{\zeta} \frac{\partial f}{\partial x_i}, \hat{\psi} \rangle$$

Since $S_f[\psi] = \text{Re} \langle (S_f)(\psi), \psi \rangle \propto \text{Re} \langle (\hat{S}_f)(\hat{\psi}), \hat{\psi} \rangle$, we obtain

$$\langle \hat{S}_f \rangle_i = \kappa \sqrt{\zeta} \frac{\partial f}{\partial x_i}$$

where $\kappa$ is a constant.

Let us now introduce a vector of functions $V_f = [(S_f)(1), \ldots, (S_f)(n)] \in L_2^2(\mathbb{R}^n)$. Using we obtain $(\hat{S}_f)_i = \kappa \sqrt{\zeta} \frac{\partial f}{\partial x_i}$, and therefore $\hat{V}_f = \kappa \sqrt{\zeta} \frac{\partial f}{\partial x_i}$. Thus, the expression $\|S_{\phi} - P_{t-1}S_{\phi_{t-1}}\|^2$ in the alternating scheme can be rewritten as

$$\|V_{\phi} - P_{t-1}V_{\phi_{t-1}}\|^2_{L_2(\mathbb{R}^n)} \propto \|\kappa \sqrt{\zeta} \frac{\partial \hat{\phi}_t}{\partial x} - P_{t-1} \kappa \sqrt{\zeta} \frac{\partial \hat{\phi}_{t-1}}{\partial x} \|^2_{L_2(\mathbb{R}^n)} \propto \| \frac{\partial \hat{\phi}_t}{\partial x} - P_{t-1} \frac{\partial \hat{\phi}_{t-1}}{\partial x} \|^2 \| \zeta_{<2,\hat{\zeta}}(\mathbb{R}^n)$$

The matrix $M_f$ can also be calculated from $\hat{f}$ using the following identity:

$$\langle x_i, f[x_j] \rangle = \langle x_i, \zeta * (x_j f) \rangle \propto \langle \frac{\partial f}{\partial x_i}, \zeta \frac{\partial f}{\partial x_j} \rangle = \langle \frac{\partial \hat{f}}{\partial x_i}, \frac{\partial \hat{f}}{\partial x_j} \rangle_{L_2,\hat{\zeta}}(\mathbb{R}^n)$$

Let us introduce a function $\hat{I}$ such that $\hat{I}(\hat{f}) = I(\hat{f})$. Then, we see that all steps in Scheme can be performed with $\hat{\phi}_t$ rather than with $\phi_t$. Using Algorithm one can construct a dual algorithm that deals with inverse Fourier transforms of functions, i.e. with $F^{-1}[\phi], F^{-1}[\phi]$, $M_t = \text{Re} \left( \frac{\partial F^{-1}[\phi]}{\partial x_i}, \frac{\partial F^{-1}[\phi]}{\partial x_j} \right)_{L_2,\hat{f}}(\mathbb{R}^n)$ etc. This version of the dual alternating scheme will be used for designing numerical algorithms for the Gaussian MMD-PCA and HM-MMD-PCA.

F  Proofs for Section

**Proof of Theorem** Let $X = [x_1, \ldots, x_N]$. Note that $H_f = XSX^T$ where $S = [P(x_i \cdot x_j)]_{i,j \in [N]}$. For any $U \in O(n)$ we have

$$H_{fU} = (U^T X)(P(U^T x_i \cdot U^T x_j))_{i,j \in [N]}(U^T X)^T = U^T H_f U.$$
Algorithm 3 The alternating scheme in the dual space.

\[
P_0 \leftarrow 0, \hat{\phi}_0 \leftarrow 0
\]
for \( i = 1, \cdots, T \) do
\[
\hat{\phi}_i \leftarrow \arg \min_{\phi} I(\hat{\phi}) + \bar{\lambda} ||\phi - P_{i-1} \hat{\phi}_{i-1}||_2^2 \|f\|_{L_{2,\bar{\epsilon}}(R^n)}^2
\]
Calculate \( M_i = \left[ \text{Re} \langle \frac{\partial \phi_{i+1}}{\partial x}, \frac{\partial \phi_{i+1}}{\partial x_j} \rangle \right]_{L_{2,\bar{\epsilon}}(R^n)} \)
Find \( \{v_i\}^n_{i=1} \) s.t. \( M_i v_i = \lambda_i v_i, \lambda_1 \geq \cdots \geq \lambda_n \)
\( P_i \leftarrow \sum_{i=1}^k v_i v_i^T \)
end for
Output: \( v_1, \cdots, v_k \)

Let \( U = \{u_1, \cdots, u_n\} \) where \( \{u_i\}^n_{i=1} \) are eigenvectors such that \( H_f u_i = \lambda_i u_i \). Then, the rotated distribution \( U \) is diagonal. Note that \( \inf_{\phi \in \mathcal{P}_k} \|f_U - T\|_K = \inf_{\phi \in \mathcal{P}_k} \|f - T\|_K \).

Therefore, w.l.o.g. we can assume that principal components of \( H_f \) are \( e_1, \cdots, e_n \) and \( H_f e_i = \lambda_i e_i, i \in [n] \), where \( \{e_i\}^n_{i=1} \) is a canonical basis in \( \mathbb{R}^k \) and \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \) are eigenvalues of \( H_f \). From the latter we conclude that
\[
\langle x_i f | H | x_j \rangle = \lambda_{ij}.
\]
Using \( P(x \cdot y) = \sum_{j=0}^{l-1} c_j \sum_{\alpha \in \mathbb{N} \cup \{0\}^n} |\alpha| = j \frac{j!}{\alpha!} \lambda^n \alpha \lambda^{\alpha |x| y} \), we obtain
\[
\langle x_i f | H | x_j \rangle = \sum_{j=0}^{l-1} c_j \sum_{\alpha \in \mathbb{N} \cup \{0\}^n} |\alpha| = j \frac{j!}{\alpha!} (\lambda^n \alpha \lambda^{\alpha})^2 = \lambda_i. \quad (109)
\]

For an input distribution \( f(x) = \frac{1}{N} \sum_{i=1}^N \delta^n(x - x_i) \), let us denote \( f'(x) = \frac{1}{N} \sum_{i=1}^N \delta^n(x_i) \lambda\delta^n(x - x_i) \), \( \alpha \) \( s \) and \( t \) \( s \) equals the first \( k \) components of \( z \in \mathbb{R}^n \). By construction,
\[
\lambda_i \sum_{j=0}^{l-1} c_j \sum_{\alpha \in \mathbb{N} \cup \{0\}^n} |\alpha| = j \frac{j!}{\alpha!} (\lambda^n \alpha \lambda^{\alpha})^2 = \lambda_i. \quad (110)
\]
whenever \( i \in [n] \setminus [k] \) for at least one \( j \in [s] \). Therefore,
\[
\|f - f'|_K^2 = \sum_{i=1}^n \sum_{j=0}^{l-1} c_j \sum_{\alpha \in \mathbb{N} \cup \{0\}^n} |\alpha| = j \frac{j!}{\alpha!} (\lambda^n \alpha \lambda^{\alpha})^2 + \lambda_i. \quad (111)
\]

The second sum \( F_2 \) equals \( \sum_{i=k+1}^n \lambda_i \). Let us compare the first sum, \( F_1 \), with the second, \( F_2 \). \( F_1 \) is a sum of positive factors of \( (\lambda^n \alpha \lambda^{\alpha})^2 \) where \( \alpha_k \neq 0, \alpha_{k+1:n} \neq 0 \). Let \( e_i \in \{\mathbb{N} \cup \{0\}\}^n \) be an \( i \)th canonical unit vector and \( \alpha_i \) denote an \( i \)th component of \( \alpha \). The coefficient in front of \( (\lambda^n \alpha \lambda^{\alpha})^2 \) in \( F_1 \) equals \( A_\alpha = c_{|\alpha|-1}(\alpha - 1)! \sum_{i=k+1}^n \frac{1}{(\alpha - e_i)!} = \frac{\epsilon_{|\alpha|-1}(\alpha - 1)!}{\alpha!} |\alpha_k| \) and the coefficient in front of \( (\lambda^n \alpha \lambda^{\alpha})^2 \) in \( F_2 \) equals \( B_\alpha = c_{|\alpha|-1}(\alpha - 1)! \sum_{i=k+1}^n \frac{1}{(\alpha - e_i)!} = \frac{\epsilon_{|\alpha|-1}(\alpha - 1)!}{\alpha!} |\alpha| \). Since \( |\alpha_k| \leq 1 \) and \( |\alpha_{k+1:n}| \leq l - 1 \), we conclude that \( A_\alpha \leq (l - 1)B_\alpha \). Therefore, \( F_1 \leq (l - 1)F_2 \).

Thus, overall we have
\[
\|f - f'|_K^2 \leq \sum_{i=k+1}^n \lambda_i. \quad (111)
\]

From \( \|T_{\mu} - T_{\nu}\|_K^2 = d_{\text{MMD}}(\mu, \nu)^2 \) the statement of theorem directly follows.

In our proof of Theorem 11, we will need the following classical theorem.
Theorem 14 (The Gaussian Poincaré inequality). Let \( g : \mathbb{R}^n \rightarrow \mathbb{R} \) be a smooth function, then
\[
\text{Var}_{x \sim N(0, \sigma^2 I_n)}[g(x)] \leq \sigma^2 \mathbb{E}_{x \sim N(0, \sigma^2 I_n)}\| \nabla g(x) \|^2.
\]
(114)

Proof of Theorem\[\] Let us assume w.l.o.g. that principal components of \( H_f \) are \( e_1, \cdots, e_n \) and \( H_f e_i = \lambda_i e_i, \ i \in [n] \), where \( \{e_i\}_{i=1}^n \) is a canonical basis in \( \mathbb{R}^n \) and \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \) are eigenvalues of \( H_f \). From the latter we conclude that \( \langle x_i f \rangle H [x_j f] = \lambda_j \delta_{ij} \).

Note that \( H(x, y) = F^{-1}[G_\sigma](x - y) \). Let \( \hat{f} = F[f] \),
\[
\hat{f}'(x) = \mathbb{E}_{y \sim N(0, \sigma^2 I_{n-k})}[\hat{f}(x_1:k, y_1:n-k)]
\]
and \( f' = F^{-1}[\hat{f}'] \) where \( x = [x_1:k, x_{k+1:n}] \). From the isometry of the Fourier transform, we have
\[
\| f - f' \|^2_K = C_n \mathbb{E}_{x \sim N(0, \sigma^2 I_n)}\| \nabla_x \hat{f} - \nabla_x \hat{f}' \|^2.
\]
(115)

The latter expression decomposes into two terms. The first is
\[
\mathbb{E}_{x \sim N(0, \sigma^2 I_n)}\| \nabla_x \hat{f}'(x) \|^2 = \mathbb{E}_{x \sim N(0, \sigma^2 I_n)}\| \nabla_x \hat{f}(x) \|^2 - \mathbb{E}_{x \sim N(0, \sigma^2 I_n)}\| \nabla_x \hat{f}'(x) \|^2 = 
\mathbb{E}_{x \sim N(0, \sigma^2 I_n)}\sum_{i=k+1}^n \left( \frac{\partial \hat{f}(x)}{\partial x_i} \right)^2 = \frac{1}{C_n} \sum_{i=k+1}^n \lambda_i
\]
(116)

The second is
\[
\mathbb{E}_{x \sim N(0, \sigma^2 I_n)}\| \nabla_x \hat{f}(x) \|^2 - \mathbb{E}_{x \sim N(0, \sigma^2 I_n)}\| \nabla_x \hat{f}'(x) \|^2 = 
\mathbb{E}_{x \sim N(0, \sigma^2 I_n)}\sum_{i=1}^k \text{Var}_{x_{k+1:n} \sim N(0, \sigma^2 I_{n-k})} \left[ \frac{\partial \hat{f}(x_1:k, x_{k+1:n})}{\partial x_i} \right].
\]
(117)

The latter can be bounded using the Gaussian Poincaré inequality by
\[
\sigma^2 \mathbb{E}_{x \sim N(0, \sigma^2 I_n)} \sum_{i=1}^k \sum_{j=k+1}^n \left( \frac{\partial \hat{f}(x_1:k, x_{k+1:n})}{\partial x_i} \right)^2.
\]
(118)

After changing an order of summations one can bound the internal sum using integration by parts, i.e.
\[
\sum_{i=1}^k \int_{\mathbb{R}^n} \left| \frac{\partial \hat{f}(x)}{\partial x_j} \right|^2 G_\sigma(x) dx = \sum_{i=1}^k - \int_{\mathbb{R}^n} \frac{\partial \hat{f}(x)^*}{\partial x_j} \frac{\partial^2 \hat{f}(x)}{\partial x_i \partial x_j} G_\sigma(x) dx = 
\sum_{i=1}^k \int_{\mathbb{R}^n} \frac{\partial \hat{f}(x)^*}{\partial x_j} \left[ \frac{\partial^2 \hat{f}(x)}{\partial x_i \partial x_j} - \frac{\partial \hat{f}(x)}{\partial x_i} \right] G_\sigma(x) dx = 
\int_{\mathbb{R}^n} \frac{\partial \hat{f}(x)^*}{\partial x_j} \left( - \frac{1}{\sigma^2} \frac{\partial (x_1:k \cdot \nabla_1:k \hat{f}(x))}{\partial x_j} \right) G_\sigma(x) dx
\]
(119)

Then, using the Cauchy–Schwarz inequality we bound the latter by
\[
\left( \int_{\mathbb{R}^n} \left| \frac{\partial \hat{f}(x)}{\partial x_j} \right|^2 G_\sigma(x) dx \right)^{1/2} \left( \int_{\mathbb{R}^n} \left| \frac{\partial (x_1:k \cdot \nabla_1:k \hat{f}(x))}{\partial x_j} \right|^2 G_\sigma(x) dx \right)^{1/2}
\]
(120)

The first term equals \( \left( \frac{1}{\sigma^2} \lambda_j \right)^{1/2} \) and the second term, after making the inverse Fourier transform, is bounded by
\[
C_n^{-1/2} \left( \langle x_1:k \cdot \nabla_1:k f \rangle H \langle x_1:k \cdot \nabla_1:k f \rangle \right)^{1/2} + C_n^{-1/2} \sum_{i=1}^k \langle (x_i x_j f \mid T_i (x - y) \rangle, \langle y_i y_j f \rangle \rangle^{1/2}
\]
(121)

where \( T_i = F^{-1}[\frac{x_i^2 G_\sigma}{\sigma^2}] = e^{-\frac{x_i^2}{2}}(1 - \sigma^2 x_i^2) \). Since \( T_i (x - y) = H(x, y) (1 - \sigma^2 x_i^2 - \sigma^2 y_i^2 + 2\sigma^2 x_i y_i) \), we have \( \langle x_i x_j f \mid T_i (x - y) \rangle, \langle y_i y_j f \rangle \rangle = \langle x_i x_j f \rangle H \langle x_i x_j f \rangle - \sigma^2 \langle x_i x_j f \rangle H \langle x_i x_j f \rangle - 2\sigma^2 \langle x_i x_j f \rangle H \langle x_i x_j f \rangle + 2\sigma^2 \langle x_i x_j f \rangle H \langle x_i x_j f \rangle \). After noting that
\[
|\langle x^2 f \rangle H |x^2 f \rangle | \leq \int |x^2 f(x)| dx \cdot \int |x^2 f(x)| dx,
\]
(122)
we finally obtain
\[
\|f - f'\|_{K}^2 \leq \sum_{j=k+1}^{n} \lambda_j^{1/2} \sum_{i=1}^{k} \left( \int |x_i x_j f(x)| \, dx \right) + \\
\sqrt{2} \sigma \sqrt{\int |x_i^2 x_j f(x)| \, dx \int |x_i x_j f(x)| \, dx} + \sqrt{2} \sigma \int |x_i^2 x_j f(x)| \, dx \leq M \sum_{j=k+1}^{n} \lambda_j^{1/2}
\]
(123)
where \( M = O(\int_{R^n} \|x\|^2 \, dx + \sqrt{2} \sigma \sqrt{\int_{R^n} \|x\|^4 \, dx \int_{R^n} \|x\|^2 \, dx} + \sqrt{2} \sigma \int_{R^n} \|x\|^3 \, dx) \).

By construction, \( f' \in G_k \) and it can be approached by elements of \( G_k \) w.r.t. norm \( \| \cdot \|_{K} \). The statement of theorem directly follows from this observation.

\[\Box\]

G  A numerical alternating scheme for the Gaussian MMD-PCA

G.1 Structure of \( \mathcal{F}^{-1}[\mathcal{P}_k] \)

From theorems [1] and [2], \( \mathcal{F}^{-1}[\mathcal{P}_k] \subseteq \mathcal{F}^{+}_k \). In fact, Bochner’s theorem [50] gives us that the inverse Fourier transform of any positive finite Borel measure is a continuous positive definite function. That is, if \( f \in \mathcal{F}^{-1}[\mathcal{P}] \), then for any distinct \( y_1, \ldots, y_s \in R^n \) the matrix \( \{f(y_i - y_j)\}_{i,j=1}^{s, s} \) is positive semidefinite. Since \( \mu(R^n) = 1 \), we additionally have \( f(0) = 1 \). Let PDF denote the set of all continuous positive definite functions on \( R^n \) and
\[
\mathcal{M}_k = \{ f \in \text{PDF} \mid \exists v_1, \ldots, v_k \in R^n, g : R^k \to C \text{ s.t. } f(x) = g(v_1^T x, \ldots, v_k^T x), f(0) = 1 \}.
\]
(124)
Thus, the following characterization of \( \mathcal{F}^{-1}[\mathcal{P}_k] \) becomes evident.

\textbf{Theorem 15.} \( \mathcal{F}^{-1}[\mathcal{P}_k] = \mathcal{M}_k \).

G.2 The dual form of the Gaussian MMD-PCA

Recall that \( k(x) = G_k^* (x) \). Let us define another Gaussian kernel \( \gamma(x) = e^{-\|x\|^2/2} = \mathcal{F}^{-1}[k] \). Let \( p_{\text{data}}(x) \) denote the characteristic function of the random vector \( X_{\text{data}} \sim \mu_{\text{data}} \). By definition, \( p_{\text{data}}(x) = \mathbb{E}[e^{ix X_{\text{data}}^T x}] = \frac{1}{N} \sum_{i=1}^{N} e^{ix_i^T x} \).

Thus, \( p_{\text{data}} = \mathcal{F}^{-1}[\mu_{\text{data}}] \) and \( \mu_{\text{data}} = \mathcal{F}[p_{\text{data}}] \).

Using the isometry property of the inverse Fourier transform for \( L_2(R^n) \) and the convolution theorem, we see that
\[
d_{\text{MMD}}(\mu, \nu) = \|k * \mu - k * \nu\|_{L_2(R^n)} \propto \|\gamma(x)(\mathcal{F}^{-1}[\mu](x) - \mathcal{F}^{-1}[\nu](x))\|_{L_2(R^n)}.
\]
(125)
Thus, from Theorem 15 we obtain that the task [17] is equivalent to
\[
\|p_{\text{data}} - q\|_{L_2, \gamma^2 (R^n)} \to \min_{q \in \mathcal{M}_k}
\]
(126)
where \( L_2, \gamma^2 (R^n) \) with \( d\nu = \gamma^2 dx \).

G.3 Algorithms for the Gaussian MMD-PCA

Let \( \Pi_k : G_k \to \{1, +\infty\} \) and \( M_k : \mathcal{F}_k \to \{1, +\infty\} \) be simple penalty functions:
\[
\Pi_k(\phi) = 1 \text{ if } \phi \in \mathcal{P}_k \text{ and } \Pi_k(\phi) = \infty \text{ otherwise}
\]
\[
M_k(\phi) = 1 \text{ if } \phi \in \mathcal{M}_k \text{ and } M_k(\phi) = \infty \text{ otherwise}
\]
Then, the task [17] is equivalent to
\[
I(\phi) = d_{\text{MMD}}^2(\mu_{\text{data}}, \phi)\Pi_k(\phi) \to \inf_{\phi \in G_k}.
\]
(127)
From the result of the previous section we see that if \( I(\phi) = \tilde{I}(\phi) \), then
\[
\tilde{I}(\phi) = \|p_{\text{data}} - \phi\|^2_{L_2, \gamma^2 (R^n)} M_k(\phi).
\]
(128)
Thus, the Algorithm [4] is an adaptation of Algorithm [5] to MMD-PCA.
Algorithm 4 The alternating scheme in the dual space for the Gaussian MMD-PCA

\begin{align*}
P_0 & \leftarrow 0, q_0 \leftarrow 0 \\
\text{for } t = 1, \ldots, T & \text{ do} \\
1 \qquad q_t & \leftarrow \arg\min_{q \in M_k} \int_{\mathbb{R}^n} \gamma(x)^2 |p_{\text{data}}(x) - q(x)|^2 dx + \lambda \int_{\mathbb{R}^n} \hat{\zeta}(x) \left\| \frac{\partial q_t}{\partial x} - P_{t-1} \frac{\partial q_{t-1}}{\partial x} \right\|^2_2 dx \\
2 \quad \text{Calculate } M_t & = \left[ \frac{\partial q_t}{\partial x_1}, \ldots, \frac{\partial q_t}{\partial x_n} \right]_{L_2(\mathbb{R}^n)} \\
3 \quad \text{Find } \{v_i\}_i^m \text{ s.t. } M_t \hat{\lambda} v_i = \lambda_i v_i, \lambda_1 \geq \cdots \geq \lambda_n \\
4 \quad P_t & \leftarrow \sum_{i=1}^k v_i v_i^T \\
\text{end for} \\
\text{Output: } \mathcal{L} = \text{span}(v_1, \ldots, v_k)
\end{align*}

If the function \( p_{\text{data}} \) is real-valued, then only real-valued functions can appear in the Algorithm 4. This assumption can be satisfied by adding reflections of initial points to the dataset (after it was centered).

At step 1, we search over \( q \) given in the following parameterized form:

\[ q_\theta(x) = \sum_{i=1}^{\text{nn}} \alpha_i \cos(\omega_i^T x) \]  

where \( \alpha_i > 0 \) and \( \sum_{i=1}^{\text{nn}} \alpha_i = 1 \). In our implementation, we set \( [\alpha_i]_{i=1}^{\text{nn}} = \text{softmax}([u_i]_{i=1}^{\text{nn}}) \) and \( u_i \)'s are unconstrained. The number of neurons in a single layer neural network with a cosine activation function, \( \text{nn} \), is a hyperparameter. Let us denote parameters \( \{\omega_i, u_i\}_{i=1}^{\text{nn}} \) by \( \theta \). It is easy to see the function \( q_\theta \) is positive definite. Moreover, using Theorem 2 from [51], it can be shown that a set of all such functions, i.e. the convex hull of \( \{\cos(\omega^T x) | \omega \in \mathbb{R}^n\} \), is dense in a set of real-valued functions from \( M_k \). Though this parameterization is quite natural, finding architectures with more expressive power in a space of real-valued positive definite functions is an open problem.

Now, to minimize

\[ \Psi(\theta) = \int_{\mathbb{R}^n} \gamma(x)^2 |p_{\text{data}}(x) - q_\theta(x)|^2 dx + \lambda \int_{\mathbb{R}^n} \hat{\zeta}(x) \left\| \frac{\partial q_\theta}{\partial x} - P_{t-1} \frac{\partial q_{t-1}}{\partial x} \right\|^2_2 dx \]

with stochastic gradient descent methods (in our case, the Adam optimizer) we need to have an unbiased estimator of

\[ \nabla_\theta \Psi(\theta) \propto \mathbb{E}_{z \sim \gamma^2} \nabla_\theta |p_{\text{data}}(z) - q_\theta(z)|^2 + \lambda \mathbb{E}_{z' \sim \hat{\zeta}} \nabla_\theta \left\| \frac{\partial q_\theta}{\partial x} (z') - P_{t-1} \frac{\partial q_{t-1}}{\partial x} (z') \right\|^2_2 \]

where \( z \sim f \) denotes that the random vector \( z \) is sampled according to the probability density function \( \frac{f(x)}{\int_{\mathbb{R}^n} f(x) dx} \).

Thus, a natural estimator of the gradient is

\[ \frac{1}{m} \sum_{i=1}^m \nabla_\theta |p_{\text{data}}(z_i) - q_\theta(z_i)|^2 + \frac{\lambda}{m} \sum_{i=1}^m \nabla_\theta \left\| \frac{\partial q_\theta}{\partial x}(\xi_i) \right\|^2_2 - P_{t-1} \frac{\partial q_{t-1}}{\partial x}(\xi_i) \right\|^2_2 \]

where \( \{z_i\}_{i=1}^m \sim \text{iid } \gamma^2 \) and \( \{\xi_i\}_{i=1}^m \sim \text{iid } \hat{\zeta} \).

The last important issue with the practical numerical algorithm is the calculation of \( M_t \) at step 2. By construction,

\[ M_t = \mathbb{E}_{x \sim \hat{\zeta}} \frac{\partial q_t}{\partial x}(x) \frac{\partial q_t}{\partial x}(x)^T. \]

In practice we sample \( x_1, \ldots, x_l \sim \hat{\zeta} \) and estimate \( M_t \) as

\[ M_t \approx \frac{1}{l} \sum_{i=1}^l \frac{\partial q_t}{\partial x}(x_i) \frac{\partial q_t}{\partial x}(x_i)^T. \]

The details of the numerical algorithm are given below. In all our experiments with MMD-PCA we set \( \hat{\zeta} = \gamma^2 \).
Algorithm 5 The numerical algorithm for the Gaussian MMD-PCA. Hyperparameters: $\lambda, h, \sigma, l, \alpha, \beta_1, \beta_2, \text{nn}$.

\[
P_0 \leftarrow 0, \theta_0 \leftarrow 0
\]

for $t = 1, \cdots, T$ do

while $\theta$ has not converged do

Sample $\{z_i\}_{i=1}^m \sim \text{iid} \, \gamma$

Sample $\{\xi_i\}_{i=1}^m \sim \text{iid} \, \zeta$

\[
L \leftarrow \frac{1}{m} \sum_{i=1}^m |p_{\text{data}}(z_i) - q_\theta(z_i)|^2 + \lambda \frac{1}{m} \sum_{i=1}^m \| \frac{\partial q_\theta(\xi_i)}{\partial x} - P_{t-1} \frac{\partial q_{\theta_{t-1}}(\xi_i)}{\partial x} \|^2_2
\]

$\theta \leftarrow \text{Adam}(\nabla_L L, \theta, \alpha, \beta_1, \beta_2)$

end while

$\theta_t \leftarrow \theta$

Sample $\{\chi_i\}_{i=1}^l \sim \text{iid} \, \zeta$

Calculate $M_t = \frac{1}{l} \sum_{i=1}^l \frac{\partial q_\theta(\chi_i)}{\partial x} \frac{\partial q_\theta(\chi_i)}{\partial x}^T$

Find $\{v_i\}_{i=1}^l$ s.t. $M_t v_i = \lambda_i v_i$, $\lambda_1 \geq \cdots \geq \lambda_n$

$P_t \leftarrow \sum_{i=1}^k v_i v_i^T$

end for

Output: $v_1, \cdots, v_k$

H A numerical alternating scheme for HM-MMD-PCA

H.1 The dual form of HM-MMD-PCA

Due to a well-known relationship between moments of the probability measure $\mu$ and its characteristic function $p$, i.e.

\[
1^p m_{i_1, \cdots, i_s} = \frac{\partial^s p(0)}{\partial x_{i_1} \cdots \partial x_{i_s}}
\]

the task (20) is equivalent to

\[
\sum_{s=1}^4 \frac{\lambda_s}{n^s} \sum_{1 \leq i_1, \cdots, i_s \leq n} \left| \frac{\partial^s p_{\text{data}}(0)}{\partial x_{i_1} \cdots \partial x_{i_s}} - \frac{\partial^s q(0)}{\partial x_{i_1} \cdots \partial x_{i_s}} \right|^2 \rightarrow \min_{\phi \in \mathcal{M}_k}.
\]

(135)

Note that the maximum mean discrepancy distance for the Gaussian kernel and the distance based on higher moments are substantially different. Indeed, even if we set $h$ as a large value (which makes $\frac{1}{h} \approx 0$), the MMD distance, unlike the HM distance, neglects higher order derivatives of the characteristic functions in the neighbourhood of the origin. Moreover, from the dual form (135) it is clear that $d_{\text{HM}}(\mu_{\text{data}}, \nu)$ is a degenerate case of a weighted Sobolev norm between characteristic functions of $\mu_{\text{data}}$ and $\nu$.

H.2 Algorithms for HM-MMD-PCA

Analogously to the case of MMD-PCA we see that the task (20) is equivalent to:

\[
I(\hat{\phi}) = d_{\text{HM}}(\mu_{\text{data}}, \hat{\phi})^2 \Pi_k(\hat{\phi}) \rightarrow \inf_{\phi \in \mathcal{B}_k} I(\hat{\phi})
\]

and

\[
I(\hat{\phi}) = \sum_{s=1}^4 \frac{\lambda_s}{n^s} \sum_{1 \leq i_1, \cdots, i_s \leq n} \left| \frac{\partial^s p_{\text{data}}(0)}{\partial x_{i_1} \cdots \partial x_{i_s}} - \frac{\partial^s \hat{\phi}(0)}{\partial x_{i_1} \cdots \partial x_{i_s}} \right|^2 M_k(\hat{\phi}).
\]

(137)

Thus, the Algorithm 3 is an adaptation of Algorithm 5 to HM-MMD-PCA.

Again, as in a numerical algorithm for MMD-PCA, at step 1, we search over $q$ given in the form (129). The objective of step 1 can be represented as

\[
\Phi(\theta) = \sum_{s=1}^4 \lambda_s \mathbb{E}_{z_{i_1, \cdots, i_s} \sim \mathcal{U}(1, n)} \left( \frac{\partial^s (p_{\text{data}} - q_\theta)(0)}{\partial x_{i_1} \cdots \partial x_{i_s}} \right)^2 + \lambda \mathbb{E}_{z \sim \zeta} \left( \frac{\partial q_\theta}{\partial x}(z) - P_{t-1} \frac{\partial q_{\theta_{t-1}}}{\partial x}(z) \right)^2
\]

(138)

where $\mathcal{U}(1, n)$ is the discrete uniform distribution over $\{1, \cdots, n\}$. To apply the stochastic gradient descent methods we need to have an unbiased estimator of $\nabla_\theta \Phi(\theta)$ which is equal to

\[
\sum_{s=1}^4 \lambda_s \mathbb{E}_{z_{i_1, \cdots, i_s} \sim \mathcal{U}(1, n)} \nabla_\theta \left( \frac{\partial^s (p_{\text{data}} - q_\theta)(0)}{\partial x_{i_1} \cdots \partial x_{i_s}} \right)^2 + \lambda \mathbb{E}_{z \sim \zeta} \left( \frac{\partial q_\theta}{\partial x}(z) - P_{t-1} \frac{\partial q_{\theta_{t-1}}}{\partial x}(z) \right)^2
\]

(139)
Algorithm 6 The alternating scheme in the dual space for HM-MMD-PCA

\[ P_0 \leftarrow 0, \theta_0 \leftarrow 0 \]
for \( t = 1, \cdots, T \) do
1. \( q_t \leftarrow \arg \min_{q \in \mathbb{R}^d} \sum_{s=1}^S \frac{\lambda_s}{m_s} \sum_{i=1 \leq i_1, \cdots, i_n \leq n} \left| \frac{\partial^r \rho_{\text{data}}(0)}{\partial x_{i_1} \cdots \partial x_{i_n}} - \frac{\partial^r \rho(0)}{\partial x_{i_1} \cdots \partial x_{i_n}} \right|^2 + \lambda \int_{\mathbb{R}^n} \hat{\zeta}(x) \left( \frac{\partial \theta^T}{\partial x} - P_{t-1} \frac{\partial \theta_{t-1}}{\partial x} \right)^2 dx \]
2. Calculate \( M_t = \left[ \frac{\partial q_t}{\partial x_{i_1}}, \frac{\partial q_t}{\partial x_{i_2}} \right]_{L_2,1}^{(\mathbb{R}^n)} \)
3. Find \( \{v_i\}_{i=1}^k \) s.t. \( M_t v_i = \lambda_i v_i, \lambda_1 \geq \cdots \geq \lambda_n \)
4. \( P_t \leftarrow \sum_{i=1}^k v_i v_i^T \)
end for
Output: \( L = \text{span}(v_1, \cdots, v_k) \)

Thus, a natural estimator of the gradient is:
\[
\sum_{s=1}^S \frac{\lambda_s}{m_s} \sum_{i=1}^{m_i} \nabla \theta \left( \frac{\partial^r (\rho_{\text{data}} - \rho)_{\text{data}}}{\partial x_{i_1} \cdots \partial x_{i_n}} \right)^2 + \frac{\lambda}{m_2} \sum_{i=1}^{m_2} \nabla \theta \left( \frac{\partial \theta_{\text{data}}}{\partial x_{i_1}} \right) - P_{t-1} \frac{\partial \theta_{t-1}}{\partial x} \right)^2 \quad (140)
\]
where \( \{a[i, j]\}_{i=1, j=1}^{\sum \lambda_{s(i), j}} \sim \text{iid} \ U(1, n) \) and \( \{\xi\}_{i=1}^{n_2} \sim \text{iid} \hat{\zeta} \). Overall, we obtain the following Algorithm 7.

Algorithm 7 The numerical algorithm for HM-MMD-PCA. Hyperparameters: \( \hat{\lambda}, \{\lambda_s\}_{s=1,4} \), \( m_1, m_2, l, \alpha, \beta_1, \beta_2, \text{nn.} \)

\[ P_0 \leftarrow 0, \theta_0 \leftarrow 0 \]
for \( t = 1, \cdots, T \) do
while \( \theta \) has not converged do
1. Sample \( \{a[i, j]\}_{i=1, j=1}^{\sum \lambda_{s(i), j}} \sim \text{iid} \ U(1, n) \)
2. Sample \( \{\xi\}_{i=1}^{n_2} \sim \text{iid} \hat{\zeta} \)
3. \( L \leftarrow \sum_{i=1}^{n_2} \sum_{j=1}^{m_1} \nabla \theta \left( \frac{\partial^r (\rho_{\text{data}} - \rho)_{\text{data}}}{\partial x_{i_1} \cdots \partial x_{i_n}} \right)^2 + \frac{\lambda}{m_2} \sum_{i=1}^{m_2} \nabla \theta \left( \frac{\partial \theta_{\text{data}}}{\partial x_{i_1}} \right) - P_{t-1} \frac{\partial \theta_{t-1}}{\partial x} \right)^2 \)
4. \( \theta \leftarrow \text{Adam}(\nabla \theta L, \theta, \alpha, \beta_1, \beta_2) \)
end while

\[ \theta_t \leftarrow \theta \]
Sample \( \{x_i\}_{i=1}^{n_1} \sim \text{iid} \hat{\zeta} \)
Calculate \( M_t = \frac{1}{T} \sum_{t=1}^{T} \left( \frac{\partial \theta_{\text{data}}(\xi_i)}{\partial x} \right) \left( \frac{\partial \theta_{\text{data}}(\xi_i)}{\partial x} \right)^T \)
Find \( \{v_i\}_{i=1}^k \) s.t. \( M_t v_i = \lambda_i v_i, \lambda_1 \geq \cdots \geq \lambda_n \)
\( P_t \leftarrow \sum_{i=1}^k v_i v_i^T \)
end for
Output: \( v_1, \cdots, v_k \)

I A numerical alternating scheme for WD-PCA

Let us consider the case \( p = 1 \) and denote \( W(\mu, \nu) = W_1(\mu, \nu) \). By Theorem 12, the task is equivalent to \( \min_{\mu \in \mathcal{P}_b} W(\mu, \mu_{\text{data}}) \), or to the following task:
\[
I(\phi) \rightarrow \inf_{\phi \in \mathcal{G}} \quad (141)
\]
where \( I(T_\mu) = W(\mu, \mu_{\text{data}}) \) if \( \mu \in \mathcal{P}_b \) and \( I(\phi) = \infty \), if otherwise. The alternating scheme is designed to solve the penalty form of the problem, i.e.
\[
I(\phi) + \lambda R(\phi) \rightarrow \min_{\phi \in \mathcal{S}(\mathbb{R}^n)}, \quad (142)
\]
which is equivalent to
\[
W(\phi, \mu_{\text{data}}) + \lambda R(\phi) \rightarrow \min_{\phi \in \mathcal{S}(\mathbb{R}^n)}, \quad (143)
\]
where \( \mathcal{S}_p(\mathbb{R}^n) \subseteq \mathcal{S}(\mathbb{R}^n) \) is a set of Schwartz functions that can serve as pdfs: \( \phi(\mathcal{X}) \geq 0, \int_{\mathbb{R}^n} \phi(\mathcal{X}) d\mathcal{X} = 1 \). A numerical version of the alternating scheme requires additional specifications on a) how to minimize over \( \phi \) at step 1, and b) how to estimate \( M_{\phi_t} \).
1.1 How to minimize over φ?

In the case of WD-PCA, the minimization step of the alternating scheme makes the following:

\[
\phi_t \leftarrow \arg \min_{\phi \in \mathcal{H}} \frac{1}{2} W(\phi, \mu_{\text{data}}) + \lambda \left\| S_\phi - P_{\phi_{t-1}} S_{\phi_{t-1}} \right\|^2
\]

where \( S_f = \sqrt{\mathbb{E}_x[f(x)]} \).

For a numerical implementation of that step we need to choose some family of functions that is dense in \( \mathcal{S}_p(\mathbb{R}^n) \) (or, rich enough to approach the solution \( \mu^* \)). Following the tradition of GAN research let us assume that the family is given in the following form\(^3\):

\[
\mathcal{H} = \{ \phi_\theta | \phi_\theta(x) \text{ is pdf of random vector } g_\theta(z), z \sim p(z), \theta \in \Theta \}
\]

where \( \{g_\theta(\theta \in \Theta) \} \) is a parameterized family of smooth functions (usually, a neural network) and \( p(z) \) is some fixed distribution (usually, the Gaussian distribution). Following \cite{52}, we make the assumption\(^1\). In a numerical algorithm we need an access to a procedure that samples according to \( \phi_\theta(x) \), not the function itself.

**Assumption 1.** \( \left\| g_\theta'(z') - g_\theta(z) \right\| \leq L(\theta, z)(\|\theta' - \theta\| + \|z' - z\|) \) where

\[
\mathbb{E}_{z \sim p(z)} L(\theta, z) < +\infty.
\]

Thus, instead of solving\(^{144}\) we solve:

\[
\phi_t \leftarrow \arg \min_{\phi \in \mathcal{H}} \frac{1}{2} W(\phi, \mu_{\text{data}}) + \lambda \left\| S_\phi - P_{\phi_{t-1}} S_{\phi_{t-1}} \right\|^2,
\]

taking into account that \( \phi_{t-1} \in \mathcal{H} \).

The Kantorovich-Rubinstein duality theorem gives us that:

\[
W(\phi_\theta, \mu_{\text{data}}) = \max_{f: \|f\|_1 \leq 1} \mathbb{E}_{x \sim \mu_{\text{data}}} [f(x)] - \mathbb{E}_{z \sim p(z)} [f(g_\theta(z))],
\]

which turns\(^{144}\) into the following minimax task:

\[
\phi_t \leftarrow \arg \min_{\phi \in \mathcal{H}} \max_{f: \|f\|_1 \leq 1} \mathbb{E}_{x \sim \mu_{\text{data}}} [f(x)] - \mathbb{E}_{z \sim p(z)} [f(g_\theta(z))] + \lambda \left\| S_\phi - P_{\phi_{t-1}} S_{\phi_{t-1}} \right\|^2.
\]

In practice, we choose a family of functions \( \mathcal{L} = \{ f_w | w \in \mathcal{W} \} \) and internal maximization is made over \( w \in \mathcal{W} \) with an additional penalty term that penalizes a violation of the Lipschitz condition: \( \forall x : \|f_w\| \leq 1 \).

A family of minimax algorithms for the minimization of \( W(\phi_\theta, \mu_{\text{comp}}) \) was developed in a series of papers \cite{52, 53, 54}. The standard minimax scheme that gained popularity in GAN literature iterates two steps: a) \( n_{\text{iter}} \) times make a gradient ascent over \( w \in \mathcal{W} \), b) make a gradient descent over \( \theta \). The task\(^{149}\) can be viewed as a Wasserstein GAN with an additional regularization term \( \lambda T(\theta) \) where \( T(\theta) = \left\| S_{\phi_\theta} - P_{\phi_{t-1}} S_{\phi_{t-1}} \right\|^2 \). To adapt these algorithms to the minimization of our function, we only need to have an unbiased estimator of the gradient \( \frac{\partial T}{\partial \theta} \). This estimator is needed for the generator to make its gradient descent step. The discriminator’s part of the algorithm (in which we maximize over Lipschitz functions \( f_w \)) can be set in a standard fashion — we choose \cite{55}’s version, in which the term \( \max \{0, \| \frac{\partial T}{\partial \theta} (\xi x + (1 - \xi) g_\theta(z)) \| - 1 \}^2 \) enforces Lipschitz condition (see step (**) of the Algorithm 5).

1.2 How to estimate \( \frac{\partial T}{\partial \theta} \) and \( M_{\phi_{t-1}} \)?

Another important aspect of the numerical algorithm is the complexity of estimating the matrix \( M_{\phi_{t-1}} \) at step (**). The following theorem shows that we only need to sample \( z \sim p \) a sufficient number of times to estimate \( \frac{\partial T}{\partial \theta} \) and \( M_{\phi_{t-1}} \).

**Theorem 16.** If \( \phi_\theta \) is pdf of the random vector \( g_\theta(z), z \sim p(z) \), then

\[
\frac{\partial T}{\partial \theta} = \mathbb{E}_{z, z' \sim p} \frac{\partial \Xi(\theta, z, z')}{\partial \theta},
\]

\[
M_{\phi_{t-1}} = \mathbb{E}_{z, z' \sim p} g_\theta(z) g_\theta(z')^T M(g_\theta(z), g_\theta(z'))
\]

where

\[
\Xi(\theta, z, z') = (g_\theta(z) \cdot g_\theta(z')) M(g_\theta(z), g_\theta(z')) - 2(g_\theta(z) \cdot P_{\phi_{t-1}} g_\theta(z)) M(g_\theta(z), g_\theta(z'))
\]

and RHS is well-defined.

\(^3\)If \( \mathcal{H} \subseteq \mathcal{S}(\mathbb{R}^n) \) is not satisfied, then we can choose \( \mathcal{H}_c = \{ \phi_\theta * G^n_\epsilon | \theta \in \Theta \} \) for a very small \( \epsilon \).
To prove the theorem we need the following lemma first.

**Lemma 5.** \( \|S_{\Theta} - PS_{\Theta}\|^2 = \mathbb{E}_{x,y \sim \phi(x \cdot y)}(M(x,y) + \mathbb{E}_{x,y \sim \psi(x \cdot P_y)}M(x,y) - 2\mathbb{E}_{x \sim \phi, y \sim \psi(x \cdot P_y)}M(x,y) \).

**Proof of lemma.**

\[
\|S_{\Theta} - PS_{\Theta}\|^2 = \|\sqrt{M[x\phi(x)]} - P\sqrt{M[x\psi(x)]}\|^2 = \\
\|\sqrt{M[x\phi(x)]} - P\psi(x)\|^2 = \sum_{i=1}^n \|\sqrt{M[x\phi(x)]} - (P\psi(x))\|^2 = \\
\sum_{i=1}^n \langle x_i\phi(x)|O_M[x_i\phi(x)] \rangle + \langle (P\psi(x)|O_M[(P\psi(x)] \rangle - 2\langle (P\psi(x)|O_M[x_i\phi(x)] \rangle = \\
\mathbb{E}_{x,y \sim \phi(x \cdot y)}M(x,y) + \mathbb{E}_{x,y \sim \psi(x \cdot P_y)}M(x,y) - 2\mathbb{E}_{x \sim \phi, y \sim \psi(x \cdot P_y)}M(x,y).
\]

**Proof of Theorem 1** Using lemma 5 we have

\[
T(\theta) = \mathbb{E}_{x,y \sim \phi}(x \cdot y)M(x,y) + \mathbb{E}_{x,y \sim \phi}(x \cdot P_{i-1}y)M(x,y) - \\
2\mathbb{E}_{x \sim \phi, y \sim \phi_{i-1}}(x \cdot P_{i-1}y)M(x,y) = \\
\mathbb{E}_{x,x' \sim \phi}(g_\theta(z) \cdot g \phi(z'))M(g_\theta(z), g \phi(z')) + \\
\mathbb{E}_{x,x' \sim \phi}(g_\theta(z) \cdot P_{i-1}g \phi_{i-1}(z'))M(g_\theta(z), g \phi_{i-1}(z')) - \\
2\mathbb{E}_{x,x' \sim \phi}(g_\theta(z) \cdot P_{i-1}g \phi_{i-1}(z'))M(g_\theta(z), g \phi_{i-1}(z')).
\]

The second term does not depend on \( \theta \). Therefore, \( \frac{\partial T}{\partial \theta} = \mathbb{E}_{x,x' \sim \phi}(x,z)\Xi(\theta, z, z') \), \( \Xi(\theta, z, z') = (g_\theta(z) \cdot g \phi(z'))M(g_\theta(z), g \phi(z')) - 2(g_\theta(z) \cdot P_{i-1}g \phi_{i-1}(z'))M(g_\theta(z), g \phi_{i-1}(z')). \)

If \( \mathbb{E}_{x,x' \sim \phi}(\frac{\partial T(\theta)}{\partial \theta} x,z) \) is well-defined (the proof of sufficiency of that condition is similar to the proof of Theorem 3 from [5]), then, using Leibniz integral rule, we obtain

\[
\frac{\partial T}{\partial \theta} = \mathbb{E}_{x,x' \sim \phi}\frac{\partial \Xi(\theta, z, z')}{\partial \theta}.
\]
The fact that
\[ M_{\hat{\phi}_t} = \mathbb{E}_{x, z \sim \rho} g_\theta(z)g_\theta(z')^T M(g_\theta(z), g_\theta(z')) \] (157)
is obvious from the definition \( M_{\phi_0} = \mathbb{E}_{x,y \sim \phi_0} xy^T M(x, y) \).

I.2.1 Definition of \( \mathcal{H} \)

Specifically, for robust PCA/outlier pursuit applications, we define \( \phi_0(x) \) as a probability density function of the random vector \( a + b \), where \( a, b \) are independent and \( a \) is the \( i \)-th column of matrix \( \theta_1 \in \mathbb{R}^{\times N} \) (where \( i \sim U(1, N) \) is sampled uniformly from \( \{1, \cdots, N\} \)), \( b = g_\theta(c) \), \( c \sim N(0, I_n) \) and \( g_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is a neural network with weights \( \theta_2 \). Thus, \( \theta = (\theta_1, \theta_2) \). It can be checked that \( \mathcal{H} \), defined in this way, satisfies the Assumption [1]. We specifically introduce the random vector \( a \) here because, according to Theorem [2], the ultimate solution of the problem corresponds to \( \theta_1 = Y \) and \( b = 0 \). This guarantees that the solution is approachable from set \( \mathcal{H} \).

J A numerical alternating scheme for SDR-ORF

For a binary classification case, given a labeled dataset \( \{(x_i, y_i)\}_{i=1}^N \), \( x_i \in \mathbb{R}^n, y_i \in C, C = \{0, 1\} \) we formulate the sufficient dimension reduction problem as the minimization task
\[ J(f) = \mathbb{E}_{(x,c)\sim\mu_{\text{data}}, \epsilon\sim N(0, \sigma^2 I_n)} L(c, f(z + \epsilon)) \rightarrow \min_{f \in \mathcal{F}} \]
(158)

where \( L(c, y) = -c \log(y) - (1 - c) \log(1 - y) \).

We apply the alternating scheme in the dual space (Algorithm [3]) to this task. We set \( M(x, y) = \zeta(x - y) \), where \( \hat{\zeta} \) is a strictly positive probability density function. A numerical version of the scheme is given below (Algorithm [9]). At every iteration \( t = 1, \cdots, T \) of the Algorithm [3] we solve the task (in our case \( \hat{I} = J \))
\[ \hat{\phi}_t \leftarrow \arg \min_{\phi} \hat{I}(\phi) + \tilde{\lambda} \parallel \frac{\partial \hat{I}}{\partial x} - P_{t-1} \frac{\partial \phi_{t-1}}{\partial x} \parallel^2_{L_2, \zeta(\mathbb{R}^n)}. \] (159)

In a numerical version of the algorithm we assume that \( \hat{\phi} \) is given as a neural network \( f_\theta \), i.e. our task becomes
\[ \theta_t \leftarrow \arg \min_{\theta} J(f_\theta) + \tilde{\lambda} \mathbb{E}_{\xi \sim \zeta} \parallel \frac{\partial f_\theta}{\partial x} (\xi) - P_{t-1} \frac{\partial f_{\theta_{t-1}}}{\partial x} (\xi) \parallel^2. \] (160)

The gradient of the function \( \Phi(\theta) = J(f_\theta) + \tilde{\lambda} \mathbb{E}_{\xi \sim \zeta} \parallel \frac{\partial f_\theta}{\partial x} (\xi) - P_{t-1} \frac{\partial f_{\theta_{t-1}}}{\partial x} (\xi) \parallel^2 \) equals
\[ \frac{\partial \Phi(\theta)}{\partial \theta} = \mathbb{E}_{(x,c)\sim\mu_{\text{data}}, \epsilon\sim N(0, \sigma^2 I_n)} \frac{\partial}{\partial \theta} L(c, f_\theta(z + \epsilon)) + \tilde{\lambda} \mathbb{E}_{\xi \sim \zeta} \frac{\partial}{\partial \theta} \parallel \frac{\partial f_\theta}{\partial x} (\xi) - P_{t-1} \frac{\partial f_{\theta_{t-1}}}{\partial x} (\xi) \parallel^2. \]

That is why \( \nabla \theta L \) (given to Adam optimizer in the gradient descent loop) in the Algorithm [9] is an unbiased estimator of \( \frac{\partial \Phi(\theta)}{\partial \theta} \). Thus, in the “while loop” we find optimal \( \phi_t = f_{\theta_t} \).

According to Algorithm [3] the next goal is to estimate
\[ M_t = \mathbb{E}_{\chi \sim \hat{\zeta}} \left( \frac{\partial \hat{\phi}_t}{\partial x} (\chi) \frac{\partial \hat{\phi}_t}{\partial x} (\chi)^T \right). \]

It is easy to see that
\[ M_t = \mathbb{E}_{\chi \sim \hat{\zeta}} \frac{\partial \hat{\phi}_t}{\partial x} (\chi) \frac{\partial \hat{\phi}_t}{\partial x} (\chi)^T = \mathbb{E}_{\chi \sim \hat{\zeta}} \frac{\partial f_{\theta_t}}{\partial x} (\chi) \frac{\partial f_{\theta_t}}{\partial x} (\chi)^T. \] (162)

From the last we see that the matrix \( M_t \) can be estimated by sampling \( \chi \sim \hat{\zeta} \) a sufficient number of times (the parameter \( l \) in our algorithm). All the rest is identical to Algorithm [3].

The regression version of the algorithm can be obtained by setting \( L(c, c') = (c - c')^2 \). Implementations for different databases can be found at [github]
Algorithm 9 The numerical alternating scheme for SDR-ORF. We use $v = 1.0$, $\hat{\zeta}(x) = G_{0,\lambda}(x)$ and default values of $\lambda = 10, m \approx 50, m' = 100, l = 30000, \alpha = 0.0001, \beta_1 = 0.5, \beta_2 = 0.9$

\[ P_0 \leftarrow 0, \theta_0 \leftarrow 0 \]

for $t = 1, \cdots, T$ do

\hspace{1em} while $\theta$ has not converged do

\hspace{2em} Sample $\{(z_i, c_i)\}_{i=1}^m \sim P_{\text{data}}$

\hspace{2em} Sample $\{\epsilon_i\}_{i=1}^m \sim N(0, v^2 I_n)$

\hspace{2em} Sample $\{\xi_i\}_{i=1}^{m'} \sim \hat{\zeta}$

\hspace{2em} $L \leftarrow \frac{1}{m} \sum_{i=1}^m L(c_i, f_\theta(z_i + \epsilon_i)) + \frac{\lambda}{m'} \sum_{i=1}^{m'} \| \frac{\partial f_{\theta}(\xi_i)}{\partial x} \| - P_{t-1} \frac{\partial f_{\theta}(\xi_i)}{\partial x} \|^2$

\hspace{2em} $\theta \leftarrow \text{Adam}(\nabla_\theta L, \theta, \alpha, \beta_1, \beta_2)$

end while

\hspace{1em} Sample $\{\chi_i\}_{i=1}^l \sim \hat{\zeta}$

\hspace{1em} Calculate $M_t = \frac{1}{l} \sum_{i=1}^l \frac{\partial f_{\theta}(\chi_i)}{\partial x} \frac{\partial f_{\theta}(\chi_i)}{\partial x}^T$

\hspace{1em} Find $\{v_i\}_{i=1}^k$ s.t. $M_t v_i = \lambda_i v_i, \lambda_1 \geq \cdots \geq \lambda_n$

\hspace{1em} $P_t \leftarrow \sum_{i=1}^k v_i v_i^T$

end for

Output: $v_1, \cdots, v_k$