Journal of Mathematical Imaging and Vision (2020) 62:376–394
https://doi.org/10.1007/s10851-019-00902-2

On Orthogonal Projections for Dimension Reduction and Applications in Augmented Target Loss Functions for Learning Problems

A. Breger[1] · J. I. Orlando[2] · P. Harar[3] · M. Dörfler[1] · S. Klimscha[2] · C. Grechenig[2] · B. S. Gerendas[1,2] · U. Schmidt-Erfurth[2] · M. Ehler[1]

Received: 30 December 2018 / Accepted: 16 August 2019 / Published online: 23 October 2019

Abstract
The use of orthogonal projections on high-dimensional input and target data in learning frameworks is studied. First, we investigate the relations between two standard objectives in dimension reduction, preservation of variance and of pairwise relative distances. Investigations of their asymptotic correlation as well as numerical experiments show that a projection does usually not satisfy both objectives at once. In a standard classification problem, we determine projections on the input data that balance the objectives and compare subsequent results. Next, we extend our application of orthogonal projections to deep learning tasks and introduce a general framework of augmented target loss functions. These loss functions integrate additional information via transformations and projections of the target data. In two supervised learning problems, clinical image segmentation and music information classification, the application of our proposed augmented target loss functions increases the accuracy.

Keywords Orthogonal Projection · Dimension reduction · Preservation of data characteristics · Supervised learning · Target features

1 Introduction

Linear dimension reduction is commonly used for preprocessing of high-dimensional data in complicated learning frameworks to compress and weight important data features. In contrast to nonlinear approaches, the use of orthogonal projections is computationally cheap, since it corresponds to a simple matrix multiplication. Conventional approaches apply specific projections that preserve essential information and complexity within a more compact representation. The projector is usually selected by optimizing distinct objectives, such as information preservation of the sample variance or of pairwise relative distances. Widely used orthogonal projections for dimension reduction are variants of the principal component analysis (PCA) that maximize the variance of the projected data [37]. Preservation of relative pairwise distances asks for a near-isometric embedding, and random projections guarantee this embeddings with high probability, cf. [5,15] and see also [1,6,12,27,30,35]. The use of random projections is especially favorable for large, high-dimensional data [48], since the computational complexity is just $O(dkm)$, e.g., using the construction in [1], with $d, k \in \mathbb{N}$ being the original and lower dimensions and $m \in \mathbb{N}$ the number of samples. In contrast, PCA needs $O(d^2m) + O(d^3)$ operations [24]. Moreover, tasks that do not have all data available at once, e.g., data streaming, ask for dimension reduction methods that are independent of the data.

In the present manuscript, we study orthogonal projections regarding the interplay between

(O1) preservation of variance,
(O2) preservation of pairwise relative distances,

aiming for a sufficient lower-dimensional data representation. We shall consider the Euclidean distance exclusively since it is most widely used in applications, especially

✉ A. Breger
anna.breger@univie.ac.at

[1] Department of Mathematics, University of Vienna, Vienna, Austria

[2] Department of Ophthalmology, Medical University of Vienna, Vienna, Austria

[3] Department of Telecommunications, Brno University of Technology, Brno, Czech Republic

for error estimation. On manifolds, the geodesic distance is locally equivalent to the Euclidean distance. The two objectives (O1) and (O2) are directly addressed by PCA (O1) and random projections (O2). We achieve the following goals: First, we clarify mathematically and numerically that the two objectives are competing, i.e., PCA and random projections preserve different kinds of information. Depending on the objectives, we discuss beneficial choices of orthogonal projections and numerically find a balancing projector for a given data set. Finally, we define a general framework of augmented target (AT) loss functions for deep neural networks that integrate information about target characteristics via features and projections. We observe that our proposed methodology can increase the accuracy in two deep learning problems.

In contrast to conventional approaches, we study the joint behavior of the two objectives with respect to the entire set of orthogonal projectors. By analyzing the correlation between the variance and pairwise relative distances of projected data, we observe that (O1) and (O2) are competing and usually cannot be reached at the same time. In classification experiments with support vector machine and shallow neural networks, we investigate heuristic choices of projections applied to the input features.

In view of learning frameworks, we utilize features and projections on target data. The class of augmented target loss functions incorporates suitable transformations and projections that provide beneficial representations of the target space. It is applied in two supervised deep learning problems dealing with real-world data.

The first experiment is a clinical image segmentation problem in optical coherence tomography (OCT) data of the human retina. Related principles of dimension reduction for other clinical classification problems in OCT have already been successfully applied in [9]. In the second experiment, we aim to categorize musical instruments based on their spectrogram; see [19] for related results. Our utilized augmented target loss functions can increase the accuracy in both experiments.

The outline is as follows. In Sect. 2, we address the analysis of the competing objectives and Theorem 2.5 yields the asymptotic correlation between variance and pairwise relative distances of projected data. Section 3 prepares for the numerical investigations by recalling t-designs as considered in [10], enabling subsequent numerics. Heuristic investigations on projected input used in a straightforward classification task are presented in Sect. 4. Our framework of augmented target loss functions as modified standard loss functions for deep learning is introduced in Sect. 5. Finally, in Sects. 6 and 7 we present classification experiments on OCT images and musical instruments using aligned augmented target loss functions.

2 Dimension Reduction with Orthogonal Projections

To reduce the dimension of a high-dimensional data set $x = \{x_i\}_{i=1}^m \subset \mathbb{R}^d$, we map x onto a lower-dimensional affine linear subspace $\bar{x} + V$, where $\bar{x} := \frac{1}{m}\sum_{i=1}^m x_i$ is the sample mean and V is a k-dimensional linear subspace of $\mathbb{R}^d$ with $k < d$. This mapping is performed by an orthogonal projector $p \in \mathcal{G}_{k,d}$, where

$$\mathcal{G}_{k,d} := \{p \in \mathbb{R}^{d\times d} : p^2 = p,\ p^\top = p,\ \operatorname{rank}(p) = k\}$$

denotes the Grassmannian, so that the lower-dimensional data representation is

$$\{\bar{x} + p(x_i - \bar{x})\}_{i=1}^m \subset \bar{x} + V, \tag{2.1}$$

with $\operatorname{range}(p) = V$. A suitable choice of p within $\mathcal{G}_{k,d}$ depends on further objectives, i.e., which kind of information preservation shall be favored for subsequent analysis tasks. In the following, we consider two objectives associated with popular choices of orthogonal projectors for dimension reduction, in particular, random projectors and PCA. We will first observe that the two objectives are competing, especially in high dimensions, and then discuss consequences.

2.1 Objective (O1)

The total sample variance $\operatorname{tvar}(x)$ of $x = \{x_i\}_{i=1}^m \subset \mathbb{R}^d$ is the sum of the corrected variances along each dimension:

$$\operatorname{tvar}(x) := \frac{1}{m-1}\sum_{i=1}^m \|x_i - \bar{x}\|^2. \tag{2.2}$$

PCA aims to construct $p \in \mathcal{G}_{k,d}$, such that the total sample variance of (2.1) is maximized among all projectors in $\mathcal{G}_{k,d}$. For other equivalent optimality criteria, we refer to [49].

The total sample variance of $px = \{px_i\}_{i=1}^m \subset V$ coincides with the one of (2.1) and satisfies

$$\operatorname{tvar}(px) \le \operatorname{tvar}(x)$$

for all $p \in \mathcal{G}_{k,d}$. Thus, PCA achieves optimal variance preservation. The total variance (2.2) can also be expressed via pairwise absolute distances:

$$\operatorname{tvar}(x) = \frac{1}{m(m-1)}\sum_{i<j}\left\|x_i - x_j\right\|^2. \tag{2.3}$$

Equally, it holds that

$$\operatorname{tvar}(px) = \frac{1}{m(m-1)}\sum_{i<j}\left\|p(x_i) - p(x_j)\right\|^2, \tag{2.4}$$

which reveals that PCA maximizes the sample mean of the projected pairwise absolute distances.

2.2 Objective (O2)

In contrast to pairwise absolute distances, the Johnson–Lindenstrauss lemma targets the global property of preservation of pairwise relative distances:

Lemma 2.1 (Johnson–Lindenstrauss, cf. [15,35]). *For any* $0 < \epsilon < 1$, *any* $k \leq d, m \in \mathbb{N}$, *with*

$$\frac{4\log(m)}{\epsilon^2/2 - \epsilon^3/3} \leq k,$$

and any set $\{x_i\}_{i=1}^m \subset \mathbb{R}^d$, *there is a projector* $p \in \mathcal{G}_{k,d}$ *such that*

$$(1-\epsilon)\left\|x_i - x_j\right\|^2 \leq \tfrac{d}{k}\left\|p(x_i) - p(x_j)\right\|^2 \leq (1+\epsilon)\left\|x_i - x_j\right\|^2 \tag{2.5}$$

holds for all $i < j$.

For small $\epsilon > 0$, the projector p in Lemma 2.1 yields that all of the $\frac{m(m-1)}{2}$ pairwise relative distances

$$\left\{\frac{d}{k}\frac{\|p(x_i) - p(x_j)\|^2}{\|x_i - x_j\|^2} : i < j\right\} \tag{2.6}$$

are close to 1, i.e., the projection p preserves all scaled pairwise relative distances well. A good choice of p in Lemma 2.1 is based on random projectors[1] $P \sim \lambda_{k,d}$, where $\lambda_{k,d}$ denotes the unique orthogonally invariant probability measure on $\mathcal{G}_{k,d}$. The following theorem is essentially proved by following the lines of the proof of Lemma 2.1 in [15] after replacing the constant 4 with $(2+\tau)2$ in the respective bound on k.

Theorem 2.2 *For any* $0 < \epsilon < 1$, *any* $k \leq d, m \in \mathbb{N}$ *and any* $0 < \tau$ *with*

$$\frac{(2+\tau)2\log(m)}{\epsilon^2/2 - \epsilon^3/3} \leq k,$$

and any set $\{x_i\}_{i=1}^m \subset \mathbb{R}^d$, *the random projector* $P \sim \lambda_{k,d}$ *satisfies*

$$\left\{\frac{d}{k}\frac{\|P(x_i) - P(x_j)\|^2}{\|x_i - x_j\|^2} : i < j\right\} \in [1-\epsilon, 1+\epsilon] \tag{2.7}$$

with probability at least $1 - \frac{1}{m^\tau} + \frac{1}{m^{\tau+1}}$.

[1] We use lower case letters for samples and upper case letters for random vectors/matrices.

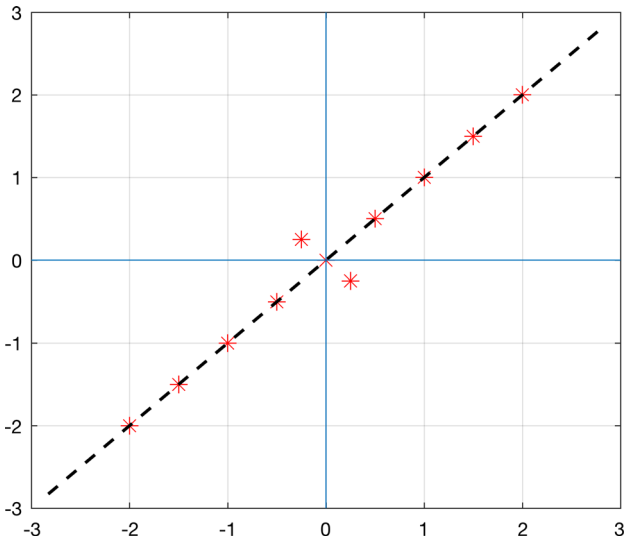


Fig. 1 A trivial example of PCA distorting smaller distances. Choosing the first principal component, PCA projects the two-dimensional data points * onto the plane of the first eigendirection (− −). The Euclidean distances of the points lying on the diagonal are preserved, whereas the two points with smaller distances are projected onto a single point (the origin)

The theorem tells that the preservation property of pairwise relative distances is achieved with high probability when choosing a random projection according to k;d. Note that the random choice is completely independent from the actual data set.

2.3 Competing Objectives

A projector p satisfying the near-isometry property (2.5) implies

$$(1-\epsilon)\tfrac{k}{d}\operatorname{tvar}(x) \leq \operatorname{tvar}(px) \leq (1+\epsilon)\tfrac{k}{d}\operatorname{tvar}(x),$$

so that the total variance of the projected data px is not preserved for $k < d$. In particular, with high probability a random projector $P \sim \lambda_{k,d}$ does not suit the objective of maximizing the total variance, and we even observe $\mathbb{E}\operatorname{tvar}(Px) = \frac{k}{d}\operatorname{tvar}(x)$; see (A.2) in the Appendix. PCA does not guarantee any local geometric property, and distances between pairs of points can be arbitrarily distorted [1]; see [39] for more robust PCA. The preservation of larger distances is favored since PCA maximizes (2.4) among all $p \in \mathcal{G}_{k,d}$ and $\|p(x_i) - p(x_j)\| \leq \|x_i - x_j\|$ holds for all $i < j$. Close but distinct points could even be projected onto a single point, which violates the preservation of pairwise relative distances; see Fig. 1.

To more quantitatively understand the relation between the two competing objectives, we consider the sample mean

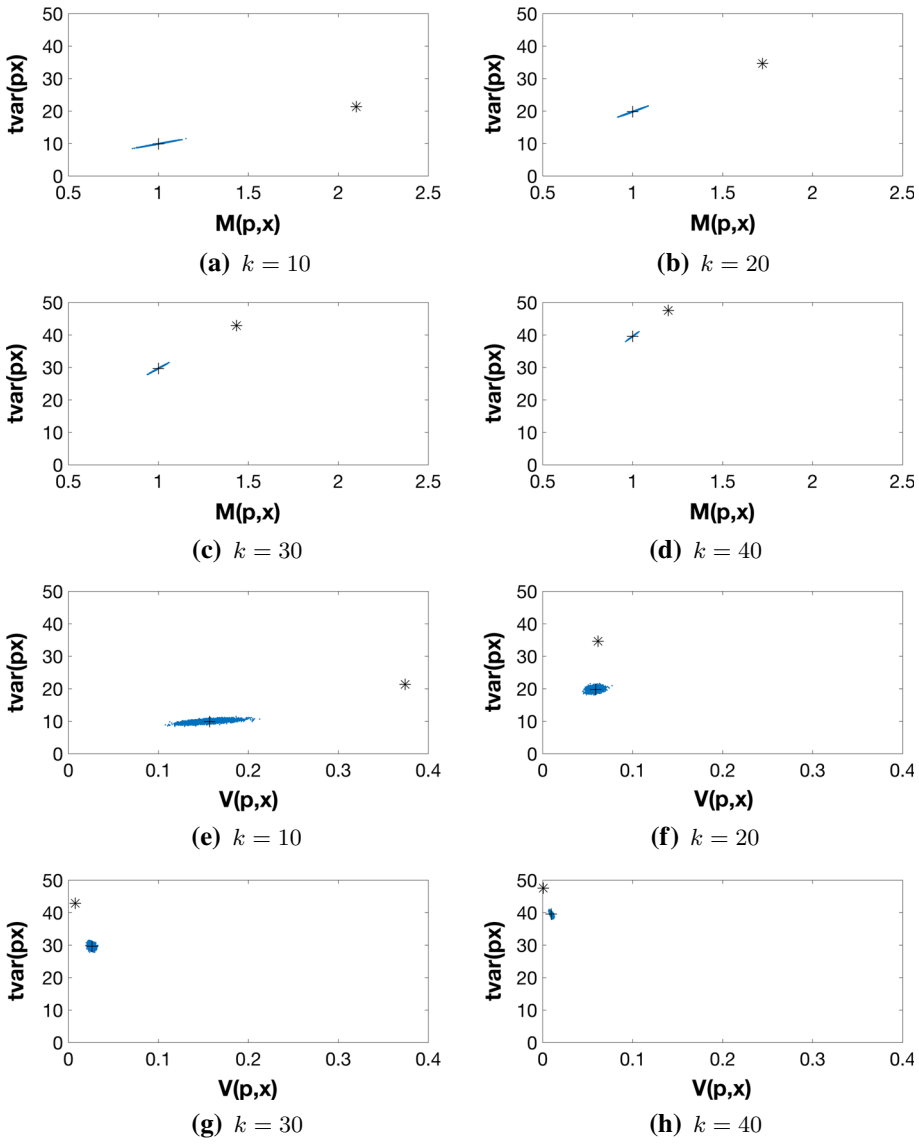


Fig. 2 Competing properties: 10,000 random projections $p \sim \lambda_{k,50}$ versus PCA ($*$), plotted concerning $\mathrm{tvar}(px)$, $\mathcal{M}(p,x)$ and $\mathcal{V}(p,x)$. The normal distributed fixed data set x has total variance $\mathrm{tvar}(x) = 49.5$. Random projections cluster around their expectation values (2.10), (2.11) and (2.12), marked by $+$

and the uncorrected sample variance of the pairwise relative distances (2.6):

$$\mathcal{M}(p,x) := \frac{2}{m(m-1)} \sum_{i<j} \frac{d}{k} \frac{\|p(x_i - x_j)\|^2}{\|x_i - x_j\|^2}, \tag{2.8}$$

$$\mathcal{V}(p,x) := \frac{2}{m(m-1)} \sum_{i<j} \frac{d^2}{k^2} \frac{\|p(x_i - x_j)\|^4}{\|x_i - x_j\|^4} - \mathcal{M}(p,x)^2. \tag{2.9}$$

Recall that good preservation of the relative pairwise distances in (2.6) asks for $\mathcal{M}(p,x)$ being close to 1 and the variance $\mathcal{V}(p,x)$ being small. In the following, we analyze $\mathrm{tvar}(px)$, $\mathcal{M}(p,x)$ and $\mathcal{V}(p,x)$ and their expectations for random $P \sim \lambda_{k,d}$.

In Fig. 2, we see a simple numerical experiment, where we first create an independent, normally distributed fixed data set $\{x_i\}_{i=1}^m$ with $x_i \in \mathbb{R}^d$ for $i = 1, \ldots, m$ and $m = 100$, $d = 50$. We then compute PCA, for $k = 10, 20, 30, 40$, as well as $n = 10{,}000$ random projections p distributed according to $\lambda_{k,50}$. In Fig. 2a–d, we can see that the more the k differs from d, the more the PCA and random projections differ concerning $\mathrm{tvar}(px)$ and $\mathcal{M}(p,x)$. Those differences may lead to diverse behavior in subsequent data analysis. Moreover, we compare $\mathcal{M}(p,x)$ and $\mathcal{V}(p,x)$ in Fig. 2e–h for the different k. We can see that again when k is much smaller than d, random projections and PCA differ more concerning the variance of pairwise distances $\mathcal{V}(p,x)$. For $k = 10$, the variance for PCA is higher in comparison with random projections (Fig. 2e); for $k = 40$ vice versa (Fig. 2h). Note that the theoretical bounds stated in Theorem 2.1 are much higher than the dimensions k used in the experiments, but the projections still preserve relative pairwise distances

very well. In [7], similar observations were made on empiric experiments with image and text data.

The amount of variance kept in the principal components comparing real-world and random data has been experimentally studied, e.g., in [29] and [46]. Both studies determine that the difference occurs mainly in the first principal component.

Remark 2.3 In the numerical example, we compare random projections and PCA directly, serving as the corresponding projections to the objectives (O1) and (O2). We observe that even for not so high-dimensional ($d = 50$) data x and $k \ll d/2$, PCA severely loses information in terms of total variance, i.e., more than 50% for $k = 10$, and more importantly loses much more information on pairwise relative distances than random projections. If both types of information are of interest, pairwise relative distances and high total variance, one should therefore favor random projections over PCA for $k \ll d/2$ to balance the two objectives (O1) and (O2) and vice versa. Note that with a large amount of data one might still want to favor random projectors since their construction is computationally much cheaper and independent from the data. On the other hand, if objective (O2) is negligible, e.g., tasks with very noisy data, then PCA would be the favorable choice for all k.

Information of data can be quantified and expressed in different ways. One crucial part in dimension reduction is the decision of what kind of information shall be kept, which depends on several parameters including the quality of the data and the analysis task. Variants of PCA, focusing on the preservation of variance, have been widely used in real-world problems with big success, especially in denoising, when the preservation of all pairwise relative distances may be counterproductive, e.g., in dMRI imaging [51] and color filter array images [56]. Drawbacks are the necessity for all data being available from the start and the high computational costs. For very high-dimensional and large data sets, the computation of PCA is often not feasible. Besides the huge benefit of data independence and low computational cost when using random projections, the near-isometry property often allows to establish that the solution found in the low-dimensional space is a good approximation to the solution in the original space [1,34].

Algorithms in machine learning often need or benefit from sufficient estimates of pairwise distances, e.g., approximate nearest-neighbor problems, supervised classification [27] and subspace clustering [26]. In [32], algorithmic applications of near-isometry embeddings have been introduced. In [7], random projections have been successfully applied to noisy and noiseless text and image data. The experimental studies include the comparison of preservation of pairwise distances between random projections and PCA. The results coincide with our observations that for $k > d/2$ PCA is able to preserve the pairwise distances sufficiently, whereas for $k < d/2$ PCA distorts them. The smaller the k, the worse the distortion, whereas random projections preserve similarities still well for very small k, while being computationally much cheaper than PCA. One should point out again that favoring preservation of pairwise distances relies on the accuracy of the original distances.

PCA and random projections are orthogonal projections favoring two different aims. We want to study in the context of the whole set of orthogonal projections if the two objectives (O1) and (O2) could be reached at the same time. We will see that the objectives act competing, and therefore we suggest a balancing projector for tasks that benefit from both objectives.

2.4 Covariances and Correlation Between Competing Objectives

For further mathematical analysis, we first introduce a more general class of probability measures on $\mathcal{G}_{k,d}$ that resemble $\lambda_{k,d}$ sufficiently well:

Definition 2.4 A Borel probability measure λ on $\mathcal{G}_{k,d}$ is called a *cubature measure of strength* t if

$$\int_{\mathcal{G}_{k,d}} f(p)\mathrm{d}\lambda_{k,d}(p) = \int_{\mathcal{G}_{k,d}} f(p)\mathrm{d}\lambda(p), \quad \text{for all } f \in \mathrm{Pol}_t(\mathbb{R}^{d^2}),$$

where $\mathrm{Pol}_t(\mathbb{R}^{d^2})$ denotes the set of multivariate polynomials of total degree t in d^2 variables.

Existence of cubature measures is studied, for instance, in [17]. For random P, we now determine the expectation values for our three quantities of interest: $\mathrm{tvar}(Px)$, $\mathcal{M}(P,x)$ and $\mathcal{V}(P,x)$. If $P \sim \lambda$ and λ is a cubature measure of strength at least 2, the identities (A.2) and (A.3) in the Appendix and a short calculation yield

$$\mathbb{E}\,\mathrm{tvar}(Px) = \tfrac{k}{d}\,\mathrm{tvar}(x), \tag{2.10}$$

$$\mathbb{E}\mathcal{M}(P,x) = 1, \tag{2.11}$$

$$\mathbb{E}\mathcal{V}(P,x) = a_{k,d}\Big(1 - \tfrac{4}{m^2(m-1)^2}\sum_{\substack{i<j\\ l<r}}\Big\langle \tfrac{x_i-x_j}{\|x_i-x_j\|}, \tfrac{x_l-x_r}{\|x_l-x_r\|}\Big\rangle^2\Big), \tag{2.12}$$

where $a_{k,d} = \frac{2d(d-k)}{k(d-1)(d+2)}$. The expected sample variance in (2.12) satisfies

$$\mathbb{E}\mathcal{V}(P,x) \le a_{k,d} \longrightarrow \frac{2}{k}, \quad \text{for} \quad d \to \infty.$$

This asymptotic bound relates to Theorem 2.2 and alludes to a near-isometry property of the type (2.7) for k sufficiently large.

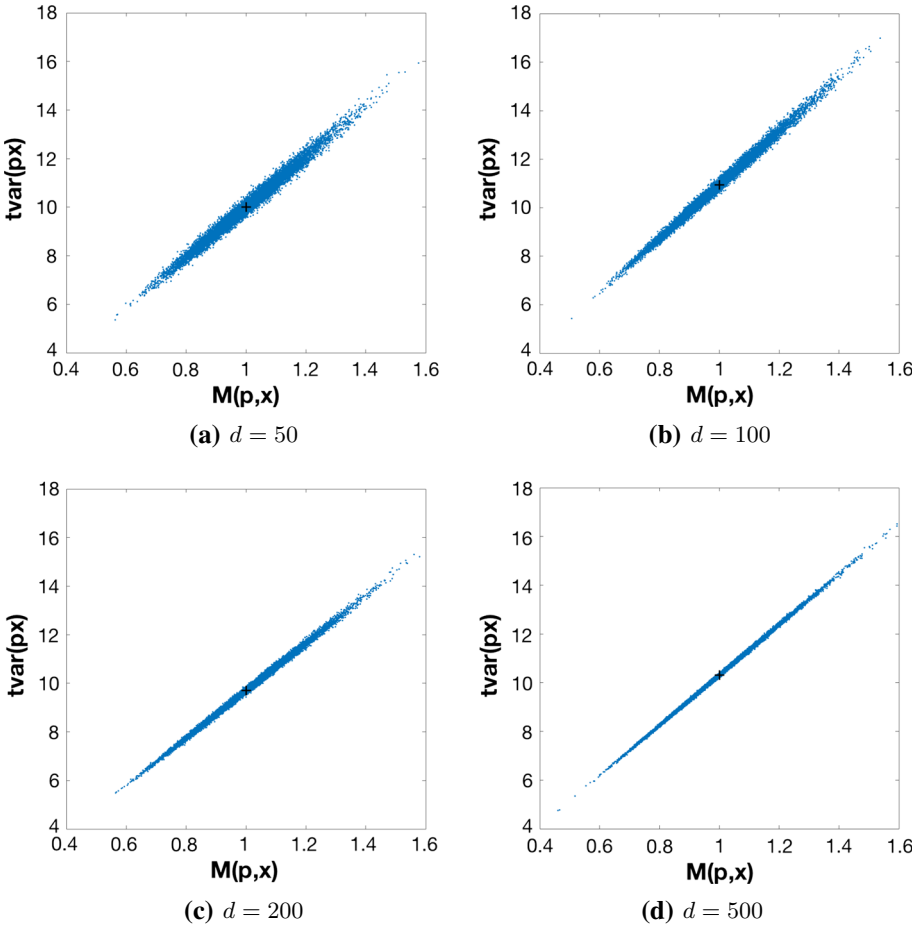


Fig. 3 For $x = \{x_i\}_{i=1}^{10} \subset \mathbb{R}^d$ with independent, normal distributed entries, we independently sample 10,000 random projectors p from $\lambda_{10,d}$ and plot $\mathcal{M}(p, x)$ versus $\mathrm{tvar}(px)$. The expectation values with respect to $P \sim \lambda$ are marked with $+$. The correlation is already 0.9916 for $d = 50$ and grows further when d increases, namely with values 0.9961, 0.9985, 0.9996 for $d = 100, 200, 500$

The following theorem provides a lower bound for random P on the population correlation

$$\mathrm{Corr}(\mathcal{M}(P,x), \mathrm{tvar}(Px)) = \frac{\mathrm{Cov}(\mathcal{M}(P,x), \mathrm{tvar}(Px))}{\sqrt{\mathrm{Var}(\mathcal{M}(P,x))}\sqrt{\mathrm{Var}(\mathrm{tvar}(Px))}}. \tag{2.13}$$

Theorem 2.5 *Let $x = \{x_i\}_{i=1}^m \subset \mathbb{R}^d$ be pairwise different and let $P \sim \lambda$, with λ being a cubature measure of strength at least 2. For $d \geq \frac{m(m-1)}{2}$, the correlation (2.13) is bounded from below by*

$$\frac{\min_{i\neq j} \|x_i - x_j\|^2}{\max_{i\neq j} \|x_i - x_j\|^2} - \frac{m(m-1)}{2d} \cdot \frac{\max_{i\neq j} \|x_i - x_j\|^2}{\min_{i\neq j} \|x_i - x_j\|^2}. \tag{2.14}$$

If $\{x_i\}_{i=1}^m \subset \mathbb{R}^d$ are random points, whose entries are independent, identically distributed with finite 4-th moments that are uniformly bounded in d, then (2.14) converges towards 1 in probability for $d \to \infty$.

The strong correlation for large dimensions d in the second part of Theorem 2.5 suggests that increasing $\mathrm{tvar}(Px)$ may also lead to increasing $\mathcal{M}(P, x)$; see Fig. 3 for illustration. Thus, large projected total variance $\mathrm{tvar}(Px)$ and the preservation of scaled pairwise distances, i.e., $\mathcal{M}(P, x)$ being close to 1, are competing properties. As discussed in Sect. 2.3, the choice of which kind of information is favorable to preserve depends on the data and the task, e.g., denoising (O1) and nearest-neighbor classification (O2). PCA and random projections are extreme in preserving either (O1) or (O2). We will heuristically study the behavior of orthogonal projections balancing both objectives in the next section and will state a numerical experiment where a balancing projector yields the highest classification accuracy.

Remark 2.6 The second part of Theorem 2.5 relates to the well-known fact that random vectors in high dimensions are almost orthogonal [4], and standard concentration of measure arguments may lead to more quantitative statements, cf. [52].

3 Preparations for Numerical Experiments

For the numerical experiments, we need finite sets of projectors that represent the overall space well, i.e., cover $\mathcal{G}_{k,d}$ properly.

3.1 Optimal Covering Sequences

Let the *covering radius* of a set $\{p_l\}_{l=1}^n \subset \mathcal{G}_{k,d}$ be denoted by

$$\varrho(\{p_l\}_{l=1}^n) := \sup_{p\in\mathcal{G}_{k,d}} \min_{1\le l\le n} \|p - p_l\|_{\mathrm{F}}, \tag{3.1}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. The smaller the covering radius, the better the set $\{p_l\}_{l=1}^n$ represents the entire space $\mathcal{G}_{k,d}$, i.e., there are smaller holes and the points $\{p_l\}_{l=1}^n$ are better distributed within $\mathcal{G}_{k,d}$. Following Lemma 2.1, we can connect finite sets of projections and their covering radius to the near-isometry property:

Lemma 3.1 *Let $\{p_l\}_{l=1}^n \subset \mathcal{G}_{k,d}$ and denote $\varrho := \varrho(\{p_l\}_{l=1}^n)$. For any $0 < \epsilon < 1$, any $m, k, d \in \mathbb{N}$ with*

$$\frac{4\log(m)}{\epsilon^2/2 - \epsilon^3/3} \le k \le d,$$

and any $\{x_i\}_{i=1}^m \subset \mathbb{R}^d$, there is $l_0 \in \{1,\dots,n\}$ such that

$$\begin{aligned}(1-\delta)\left\|x_i - x_j\right\|^2 &\le \tfrac{d}{k}\left\|p_{l_0}(x_i) - p_{l_0}(x_j)\right\|^2 \\ &\le (1+\delta)\left\|x_i - x_j\right\|^2, i < j,\end{aligned} \tag{3.2}$$

where $\delta = \epsilon + 2\varrho\sqrt{\frac{(1+\epsilon)d}{k}} + \frac{d}{k}\varrho^2$.

Proof Given an arbitrary projector $p \in \mathcal{G}_{k,d}$, there is an index $l_0 \in \{1,\dots,n\}$ such that

$$\|p_{l_0}x - px\| \le \|p_{l_0} - p\|_{\mathrm{F}}\|x\| \le \varrho\|x\|, \quad x \in \mathbb{R}^d.$$

From here, standard computations imply Lemma 3.1. We omit the details. □

The accuracy of the near-isometry property in (3.2) depends on the covering radius. Therefore, a set $\{p_l\}_{l=1}^n \in \mathcal{G}_{k,d}$ with a small covering radius ϱ is more likely to contain a projector with better preservation of pairwise relative distances. According to [11], it holds that[2] $\varrho \gtrsim n^{-\frac{1}{k(d-k)}}$ and we shall see next how to achieve this lower bound.

A set of projectors $\{p_l\}_{l=1}^n \subset \mathcal{G}_{k,d}$ is called a *t-design* if the associated normalized atomic measure $\frac{1}{n}\sum_{l=1}^n \delta_{p_l}$ is a cubature measure of strength t (see Definition 2.4); see [44] for general existence results. Any sequence of t_i-designs $\{p_l^i\}_{l=1}^{n_i} \subset \mathcal{G}_{k,d}$ with $t_i \to \infty$ satisfies

$$\varrho_i \asymp t_i^{-1}, \tag{3.3}$$

and moreover, the bound $n_i \gtrsim t_i^{k(d-k)}$ holds, cf. [11,17]. To relate n_i to ϱ_i via t_i, a sequence of t_i-designs $\{p_l^i\}_{l=1}^{n_i} \subset \mathcal{G}_{k,d}$ is called a *low-cardinality design sequence* if $t_i \to \infty$ and

$$n_i \asymp t_i^{k(d-k)}, \quad i = 1, 2, \dots. \tag{3.4}$$

For their existence and numerical constructions, we refer to [21] and [10,11]. According to [11], see also (3.3) and (3.4), any low-cardinality design sequence $\{p_l^i\}_{l=1}^{n_i}$ covers asymptotically optimal, i.e.,

$$\varrho_i \asymp n_i^{-\frac{1}{k(d-k)}}.$$

Benefiting from the covering property, we will use low-cardinality design sequences as a representation of the overall space of orthogonal projectors $\mathcal{G}_{k,d}$.

3.2 Linear Least Squares Fit

With the linear least squares fit, we can directly gain information about the relation between $\mathcal{M}(p,x)$ and $\mathrm{tvar}(px)$ for a given data set $x = \{x_i\}_{i=1}^m \subset \mathbb{R}^d$ when p varies. Given the two samples

$$\{\mathrm{tvar}(p_1x),\dots,\mathrm{tvar}(p_nx)\}, \quad \{\mathcal{M}(p_1,x),\dots,\mathcal{M}(p_n,x)\}, \tag{3.5}$$

the linear least squares fitting provides the best fitting straight line,

$$\mathrm{tvar}(p_lx) \approx s\cdot\mathcal{M}(p_l,x) + \gamma, \quad l = 1,\dots,n,$$

where s and γ are determined by the sample variances and the sample covariance. If $\{p_l\}_{l=1}^n$ is a 2-design, then the sample (co)variances coincide with the respective population (co)variances for $P \sim \lambda_{k,d}$; see Appendix A.3 for further details. It follows that

$$s = \frac{\mathrm{Cov}(\mathcal{M}(P,x), \mathrm{tvar}(Px))}{\mathrm{Var}(\mathcal{M}(P,x))} \quad \text{with } P \sim \lambda_{k,d}, \tag{3.6}$$

$$\gamma = \tfrac{k}{d}\,\mathrm{tvar}(x) - s. \tag{3.7}$$

The quantities s and γ can be directly computed, where $\mathrm{tvar}(x)$ is given by (2.2) and the covariances are stated in Corollary A.1. Note that (3.6) and (3.7) are now independent of the particular choice of $\{p_l\}_{l=1}^n$.

The correlation between the two samples (3.5) yields additional information about their relation. As before, if $\{p_l\}_{l=1}^n$ is a 2-design, then the sample correlation coincides with the population correlation (2.13) for $P \sim \lambda_{k,d}$, cf. Appendix A.3. High correlation for a specific data set x suggests that random projections and PCA preserve competing properties,

[2] We use the symbols $\lesssim$ and $\gtrsim$ to indicate that the corresponding inequalities hold up to a positive constant factor on the respective right-hand side. The notation $\asymp$ means that both relations $\lesssim$ and $\gtrsim$ hold.

whose benefits need to be assessed for the specific subsequent task.

4 Numerical Experiments in Pattern Recognition

We investigate the impact on classification accuracy when applying specific orthogonal projections to input data. The chosen real-world data yields a straightforward classification task, serving as a toy example for comparing the accuracy of several projected input data in simple learning frameworks. Projectors are chosen from a t-design in view of $\mathrm{tvar}(px)$ and $\mathcal{M}(p,x)$. For all computations made in this section, the 'Neural Network' and 'Statistics and Machine Learning' toolboxes in MATLAB R2017a are used.

We use the publicly available `iris` data set from the UCI Repository of Machine Learning Database suitable for supervised classification learning. It consists of three classes with 50 instances each, where each class refers to a type of iris plant. The instances are described by four features resulting in the input samples $\{x_i\}_{i=1}^{150} \subset \mathbb{R}^4$ and target samples $\{y_i\}_{i=1}^{150} \subset \{0,1\}^3$. For comparison, we classify the diverse input data with support vector machine (SVM) and three-layer neural networks (NN) with 5 and 10 hidden units (HU).

4.1 Choice of Orthogonal Projection

In the experiment, we use projections $p \in \mathcal{G}_{2,4}$ reducing the original dimension from $d = 4$ to $k = 2$. As a finite representation of the overall space, we use a t-design of strength 14 from a low-cardinality sequence (see Sect. 3.1) consisting of 8475 orthogonal projectors. Note that the dimension reduction in practice takes place by applying $q \in \mathcal{V}_{k,d}$ with $q^\top q = p \in \mathcal{G}_{k,d}$, where

$$\mathcal{V}_{k,d} := \{q \in \mathbb{R}^{k\times d} : qq^\top = I_k\}$$

denotes the Stiefel manifold. When taking norms, p and q are interchangeable, i.e., $\|q(x)\|^2 = \|p(x)\|^2$, for all $x \in \mathbb{R}^d$. Therefore, we can use w.l.o.g. the theory developed for p.

The projections are chosen in a deterministic manner viewing the previously described competing properties. In Fig. 4, the quantities $\mathrm{tvar}(px)$ and $\mathcal{M}(p,x)$ are pairwise plotted for all projectors in $\{p_l\}_{l=1}^{8475}$. For comparison, we choose the following projections $p \in \{p_l\}_{l=1}^{8475} \subset \mathcal{G}_{2,4}$; see Fig. 4a for a visualization.

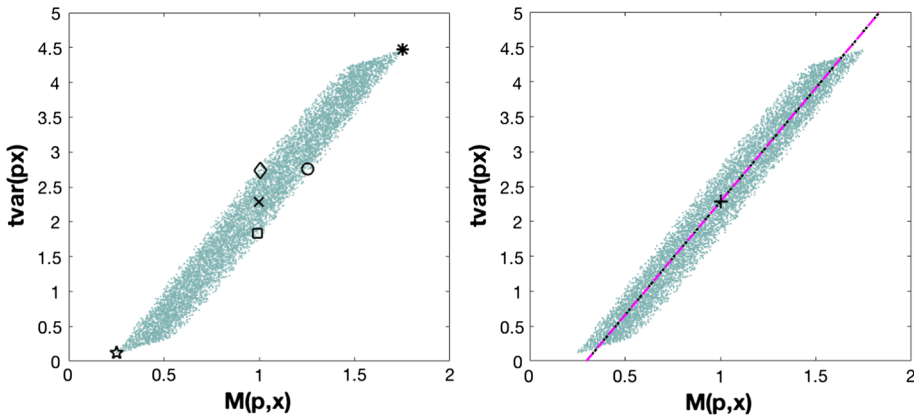
- $p_\times$ closest to the expected values 1 and $\frac{k}{d}\,\mathrm{tvar}(x)$ (see (2.10) and (2.11)),
- $p_\diamond$ preserving $\mathcal{M}(p,x) \approx 1$ and maximizing $\mathrm{tvar}(px)$,
- $p_\square$ preserving $\mathcal{M}(p,x) \approx 1$ and minimizing $\mathrm{tvar}(px)$,
- $p_\circ$ $\mathrm{tvar}(px) \approx \mathrm{tvar}(p_\diamond x)$ and maximizing $\mathcal{M}(p,x)$,
- $p_\star$ minimal $\mathrm{tvar}(px)$,
- p_* maximal $\mathrm{tvar}(px)$ (PCA).

4.2 Results

In Fig. 4b, we see the linear least squares fitting line, computed directly and via the slope and intercept as stated in (3.6) and (3.7). The correlation coefficient (2.13) is 0.98, which suggests that preserving the two properties is highly competing and needs to be balanced.

In Table 1, the classification results of the iris data are presented. We can see that in this comparison the projector $p_\diamond$, which corresponds to preserving $\mathcal{M}(p,x) \approx 1$ and maximizing $\mathrm{tvar}(px)$, yields the highest and most robust results. It even yields better results than working with the original input data. The projections that preserve $\mathcal{M}(p,x) \approx 1$ but do not take care of the magnitude of the total variance yield much worse results. On the other hand, the projections that just focus on high total variance still do not yield as high results as the projection $p_\diamond$ that balances both properties.

Remark 4.1 Given a data set x, the projector $p_\diamond$ is a good choice to balance both objectives (O1) and (O2). It can be computed by directly analyzing $\{\mathrm{tvar}(p_1 x), \ldots, \mathrm{tvar}(p_n x)\}$



(a) $\{\mathcal{M}(p_l,x), \mathrm{tvar}(p_l x)\}_{l=1}^{8475}$ **(b)** least squares fit

Fig. 4 Projections $\{p_l\}_{l=1}^{8475} \subset G_{2,4}$ from a t-design of strength 14 evaluated on the iris data set $x \subset \mathbb{R}^{4\times 150}$

Table 1 Classification results of iris data, when using projected input data in support vector machine (SVM) and shallow neural networks (NN)

Input/method	NN (10 HU)	NN (5 HU)	SVM
$\mathbf{x}$	(97.6, 1.25)	(97.5, 2.29)	(96.7, 0.15)
$\mathbf{p}_{\diamond}\mathbf{x}$	**(98.4, 0.42)**	**(98.3, 2.15)**	**(97.3, 0.06)**
$\mathbf{p}_{\times}\mathbf{x}$	(88, 1.73)	(87.9, 1.92)	(87.6, 0.63)
$\mathbf{p}_{\square}\mathbf{x}$	(87.3, 9.56)	(86.9, 10.81)	(87.7, 0.42)
$\mathbf{p}_{\circ}\mathbf{x}$	(96.8, 2.74)	(96.7, 1.77)	(96, 0.17)
$\mathbf{p}_{*}\mathbf{x}$ (PCA)	(96.9, 1.36)	(96.5, 4.07)	(96, 0.37)
$\mathbf{p}_{\star}\mathbf{x}$	(62.1, 44.30)	(58.9, 70.78)	(56, 0.61)

Mean and variance ($\times 10^{-4}$) of 1000 independent NN runs and 100 independent runs with tenfold cross-validation in SVM
Bold values indicate corresponding to the highest (= best) classification accuracy

and $\{\mathcal{M}(p_1, x), \ldots, \mathcal{M}(p_n, x)\}$ of a finite covering $\{p_l\}_{l=1}^n$ of $\mathcal{G}_{k,d}$. For higher dimensions, an accurate representation of $\mathcal{G}_{k,d}$, in order to heuristically select $p_{\diamond}$, requires large computational costs. The least squares regression line for a 2-design, as stated in 3.7, can be directly computed with low computational cost. This offers helpful information about the interplay between (O1) and (O2).

5 Augmented Target Loss Functions

In the previous section, projectors were applied to input features of shallow neural networks. In more complex architectures, such as deep neural networks, the adaption of weights can be viewed as optimization of input features, e.g., arising features can be used for transfer learning [54]. Whereas the input data are processed and optimized in each iteration, the target data stay usually unchanged during the whole learning process, serving as a measure of accuracy. The representation of the target data is one key property for successful approximation with neural networks. Here, we will introduce a general class of loss functions, i.e., augmented target (AT) loss functions, that use projections and features to yield beneficial representations of the target space, emphasizing important characteristics.

In optimization problems, additional penalty terms are used for regularization or to enforce other constraints. In deep learning, weight decay (i.e., Tikhonov regularization) is a standard adaption of the loss function to that effect. Incorporating additional underlying information via features of the output/target data has been studied in diverse settings tailored to particular imaging applications. Perceptual loss functions have been used in [31] for image super-resolution, incorporating the comparison of high-level image features that arise from pretrained convolutional neural networks, i.e., the VGG network [45]. Deep perceptual similarity metrics have been proposed in [20] for generating images, comparing image features instead of the original images. In [28], a similar approach was successfully used for style transfer and super-resolution, adding a network that defines loss functions. Anatomically constrained neural networks (ACNN) have been introduced in [40] and applied to cardiac image enhancement and segmentation. Their loss functions incorporate structural information by using autoencoders to gain features about lower-dimensional parametrization of the segmentation. Brain segmentation was studied in [22], where information about the desired structure has been added in the loss function via an adjacency matrix. It was used for fine-tuning the supervised learned network with unlabeled data, reducing the number of abnormalities in the segmentation.

The information of certain target characteristics can be very powerful and even replace the need of annotations in some tasks. In [47], label-free learning is approached by using just structural information of the desired output in the loss function instead of annotated target values.

In the following, we will define a general framework of loss functions that add information of target characteristics via features and projections in supervised learning tasks.

5.1 General Framework

Let the training data be input vectors $\{x_i\}_{i=1}^m \subset \mathbb{R}^r$ with associated target values $\{y_i\}_{i=1}^m \subset \mathbb{R}^s$. We consider training a neural network

$$f_\theta : \mathbb{R}^r \to \mathbb{R}^s,$$

where $\theta \in \mathbb{R}^N$ corresponds to the vector of all free parameters of a fixed architecture. In each optimization step for θ, the network's output $\{\hat{y}_i = f_\theta(x_i)\}_{i=1}^m \subset \mathbb{R}^s$ is compared with the targets $\{y_i\}_{i=1}^m$ via an underlying loss function L.

In contrast to ordinary learning problems with highly accurate target data, complicated learning tasks arising in many real-world problems do not yield sufficient results when optimizing neural networks with standard loss functions L, such as the widely used mean least squares error

$$L_{\mathrm{MSE}}(\{y_i\}_{i=1}^m, \{\hat{y}_i\}_{i=1}^m) := \frac{1}{m}\sum_{i=1}^m \left\| y_i - \hat{y}_i \right\|^2. \tag{5.1}$$

The training data may include important information that is obvious for humans, but poorly represented within the original target data and therefore lacks consideration in the learning process. To overcome this issue, we propose to add information tailored to the particular learning problem represented by additional features of the outputs and targets.

First, we select transformations

$$T_j : \mathbb{R}^s \to \mathbb{R}^t, \quad j = 1, \ldots, d,$$

to enable error estimation in transformed output/target spaces. Note that the transformations T_j are not required to be linear. However, they should be piecewise differentiable to enable subsequent optimization of the loss function with gradient methods. We shall allow for additional weighting of the transformations $T_1, \ldots, T_d$ to facilitate the selection of features for a specific learning problem. The previous sections suggest that orthogonal projections can provide favorable feature combinations, which essentially turns into a weighting procedure.

To enable suitable projections, we stack the d output/target features

$$T(y_i) := \begin{pmatrix} T_1(y_i)^\top \\ \vdots \\ T_d(y_i)^\top \end{pmatrix} \in \mathbb{R}^{d \times t},$$

so that applying a projector $p \in \mathcal{G}_{k,d}$ to each column of $T(y_i)$ yields $p(T(y_i)) \in \mathbb{R}^{d \times t}$. We now define the augmented target loss function with projections by

$$L_p\big(\{y_i\}, \{\hat{y}_i\}\big) := L(\{y_i\}, \{\hat{y}_i\}) + \alpha \cdot \tilde{L}\big(\{p(T(y_i))\}, \{p(T(\hat{y}_i))\}\big), \tag{5.2}$$

where $\alpha > 0$ and L and $\tilde{L}$ correspond to conventional loss functions. Apparently, L_p depends on the choice of $p \in \mathcal{G}_{k,d}$. The projection $p(T(y_i))$ weighs the previously chosen feature transformations $T(y_i)$. Standard choices of L and $\tilde{L}$ are L_{MSE}, in which case L_p becomes

$$L_p\big(\{y_i\}, \{\hat{y}_i\}\big) = \frac{1}{m} \sum_{i=1}^{m} \big\| y_i - \hat{y}_i \big\|^2 + \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} \| p(T(y_i)) - p(T(\hat{y}_i)) \|_{\mathrm{F}}^2. \tag{5.3}$$

Remark 5.1 For $k = d$, the projector p is the identity. In this case, the transformations can map onto different spaces, i.e.,

$$T_j : \mathbb{R}^s \to \mathbb{R}^{t_j}, \quad j = 1, \ldots, d,$$

and we can now write the standard augmented target loss function by

$$L_{\mathrm{AT}}\big(\{y_i\}, \{\hat{y}_i\}\big) = \sum_{j=1}^{d} \alpha_j \cdot L^j\big(\{T_j(y_i)\}, \{T_j(\hat{y}_i)\}\big), \tag{5.4}$$

where T_1 corresponds to the identity function, $L^1, \ldots, L^d$ are common loss functions and $\alpha_1, \ldots, \alpha_d > 0$ are weighting parameters.

It should be mentioned that α resembles a regularization parameter. The actual minimization of (5.1) among θ is usually performed through Tikhonov-type regularization in many standard deep neural network implementations. The formulation (5.2) adds one further variational step for beneficial output data representation.

Remark 5.2 Our proposed structure with target feature maps $T_1, \ldots, T_d$ as in (5.4) relates to multitask learning, which has been successfully used in deep neural networks [13]. It handles multiple learning problems with different outputs at the same time. In contrast to multitask learning, we aim to solve a single problem but also penalize the error in transformed spaces enhancing certain target characteristics.

For the projected feature transformations in the augmented target loss function, it is not possible to identify a balancing projection p heuristically (such as $p_\diamondsuit$ in Sect. 4), because the output y changes in each iteration when the loss function is called. In the following clinical numerical experiment we overcome this issue by choosing random projections in each optimization step and compare it to prior deterministic choices of projections, including PCA.

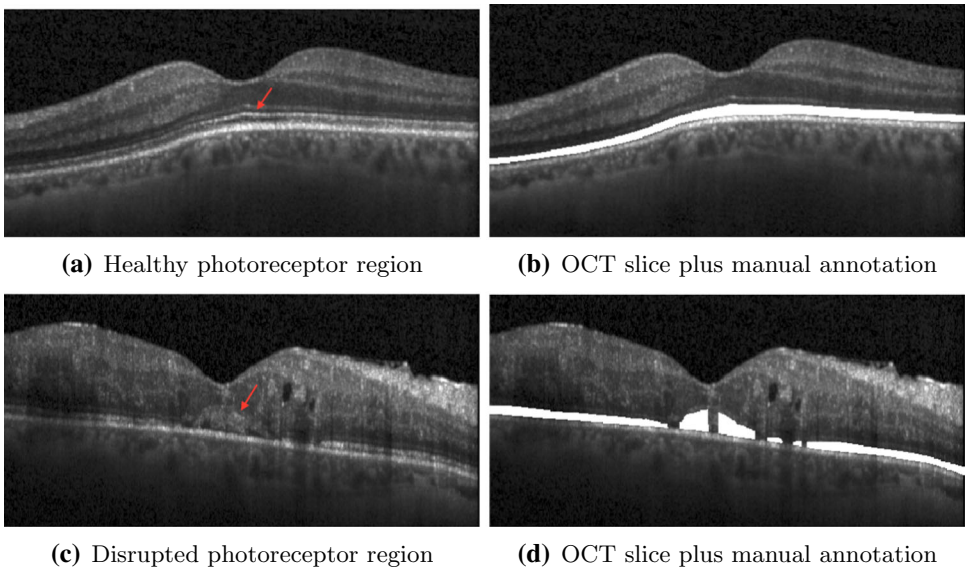
6 Application to Clinical Image Data

The first experiment is a clinical problem in retinal image analysis of the human eye, where the disruptions of the so-called photoreceptor layers need to be quantified in optical coherence tomography (OCT) images. The photoreceptors have been identified as the most important retinal biomarker for prediction of vision from OCT in various clinical publications, see e.g., [23]. As OCT technology advances, clinicians are not able to look at each slice of OCT themselves. (In mean, they get 250 slices per patient and have 3–5 minutes/patients including their clinical examination.) Therefore, automated classification of, for example, photoreceptor status is necessary for clinical guidance.

6.1 Data and Objective

In this application, OCT images of different retinal diseases (diabetic macular edema and retinal vein occlusion) were provided by the Vienna Reading Center recorded with the Spectralis OCT device (Heidelberg Engineering, Heidelberg, Germany). Each patient's OCT volume consists of 49 cross sections/slices (496 × 512 pixels) recorded in an area of 6 × 6 mm in the center of the human retina, which is the part of the retina responsible for vision. Each of the slices was manually annotated by a trained grader of the reading center. This is a challenging and time-consuming procedure that is not feasible in clinical routine but only in a research setting. The binary pixelwise annotations serve as target values, enabling a supervised learning framework (Fig. 5).

The objective is to accurately detect the photoreceptor layers and their disruptions pixelwise in each OCT slice by training a deep convolutional neural network with a suitable



(a) Healthy photoreceptor region **(b)** OCT slice plus manual annotation

(c) Disrupted photoreceptor region **(d)** OCT slice plus manual annotation

Fig. 5 OCT provides cross-sectional visualization of the human retina

loss function. The learning problem is complicated by potentially inaccurate target annotations, as studies have shown that inconsistencies between trained graders are common, cf. [50]. Moreover, the learning task is unbalanced in the sense that there are many more slices showing none or very little disruptions. We shall observe that optimization with respect to standard loss functions performs poorly in regards to detecting disruptions. The augmented target loss function proposed in the previous section can enhance the detection.

6.2 Convolutional Neural Network Learning

We implemented our experiments using Python 3.6 with Pytorch 1.0.0. A deep convolutional neural network f_θ is trained by applying the U-Net architecture reported in [43] with a sigmoid activation function and Tikhonov regularization. A set of 20 OCT volumes (980 slices) from different patients with corresponding annotations are used for training, where four volumes were used for calibration (validation set). Another two independent test volumes were identified for evaluating the results, one without any disruptions in the photoreceptor layers, whereas the other one includes a high number of disruptions.

Each OCT slice is represented by a vector $x_i \in \mathbb{R}^r$ with $r = 496 \cdot 512$. The collection $\{x_i\}_{i=1}^m$ corresponds to all slices from the training volumes, i.e., $m = 20 \cdot 49$. Further matching the notation of the previous section, we have $r = s$ and $f_\theta : \mathbb{R}^r \to \mathbb{R}^r$ with binary target vectors $y_i \in \{0, 1\}^r$. We observe that disruptions are not identified reliably when using the least squared loss function (5.1). To overcome this issues, we use the proposed augmented target loss function with least squared losses as stated in (5.3).

To enhance disruptions within the output/target space, we heuristically choose $d = 4$ local features of the original representation. They are derived from convolutions with two edge filters, T_1 (Prewitt) and T_2 (Laplacian of Gaussian), and from two Gaussian high-pass filters, yielding T_3 and T_4. Note that these feature transformations keep the same size, i.e., $T_j : \mathbb{R}^r \to \mathbb{R}^r$ for $j = 1, \ldots, d$. See Fig. 6 for example images.

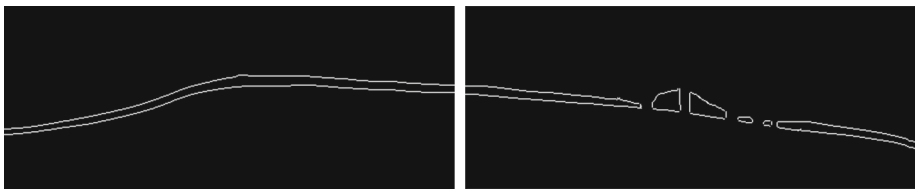
Table 2 Comparison of AUC values for photoreceptors segmentation and disruption detection

Loss function	Photoreceptors	Disruptions
L_{MSE}	0.9720	0.4399
L_p		
$p = I_4$	0.9736	0.4686
$p_{\lambda_{2,4}}$	0.9746	0.4720
p_{PCA}	0.9716	0.5331
p_{12}	**0.9755**	**0.5558**

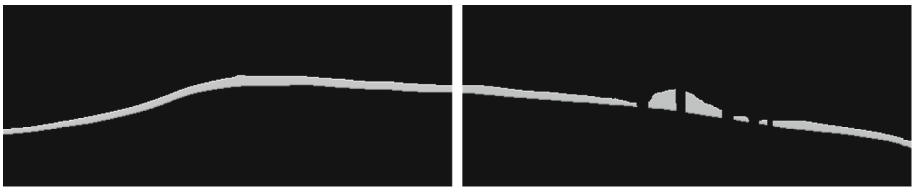
Bold values correspond to the highest AUC value, which serves us as a measure of accuracy

We can derive several augmented target loss functions L_p by choosing different $p \in \mathcal{G}_{k,d}$. In this experiment, we use the following projections:

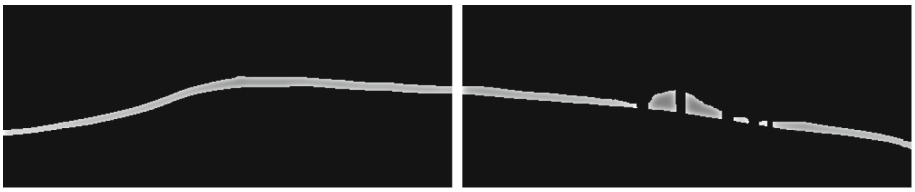
- $p = I_4$,
- $\{p_l\}_{l=1}^{15}$, all projections from a t-design of strength $2 \subset \mathcal{G}_{2,4}$ (see [10]),
- $p_{\text{PCA}} \in \mathcal{G}_{2,4}$, projection determined by PCA on the training data,
- $p_{\lambda_{2,4}}$, random projection chosen according to $\lambda_{2,4}$ in each mini-batch.



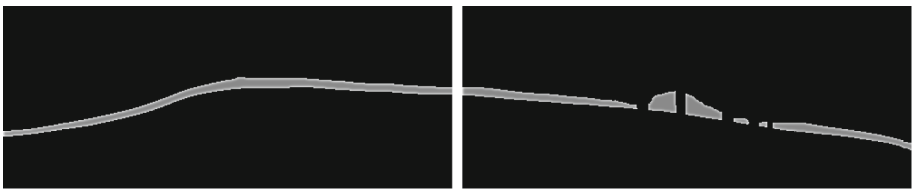
(a) Prewitt



(b) Laplacian of Gaussian (LoG)



(c) Gaussian highpass (threshold = 40)



(d) Gaussian highpass (threshold = 100)

Fig. 6 Features on output and targets that enhance edges in different ways. It is not obvious which transformations are of most importance; weighting by projections can overcome this issue

6.3 Results

Since the detection problem is highly unbalanced, we use precision/recall curves [16] for evaluating the overall performance of each loss function model. The area under the curve (AUC) was used as a numerical indicator of the success rate [41]. The higher the AUC, the better the classification.

The results of the different loss functions on the independent test set are stated in Table 2. Due to the imbalance within the data, the photoreceptor region is identified well, but disruptions are not identified reliably when using the least squared loss function (5.1). For $\alpha = 0.1$, all proposed augmented target loss functions L_p clearly increase the success rate of the disruption quantification. Note that all projections are independent from the actual data set, except PCA that was computed beforehand on the training data.

The features itself (i.e., $p = I_4$) improve the quantification, and weighting them by projections increases the results even more: using the fixed projection p_{12} from the t-design sequence $\{p_l\}_{l=1}^{15}$ on the output/target features yields the highest accuracy for photoreceptors and disruptions. This corresponds to the results of the previous sections, stating that depending on the particular data there are projections in the overall space acting beneficially. Since this projection generally cannot be found beforehand, using random projections in each loss function's evaluation step is easier, possible in practice, and independent from the data. The computation is efficient and randomization can have regularization effects that yield more robust results, cf. [34]. In the following, we will view a second classification problem based on spectrograms, where augmented target loss functions with random projections can improve the accuracy.

7 Application to Musical Data

Here, the learning task is a prototypical problem in Music Information Retrieval, namely multi-class classification of musical instruments. In analogy to the MNIST problem in image recognition, this classification problem is commonly used as a basis of comparison for innovative methods, since the ground truth is unambiguous and sufficiently many anno-

tated data are available. The input to the neural network is spectrograms of audio signals, which is the standard choice in audio machine learning. Spectrograms are calculated from the time signal using a short-time Fourier transform and taking the absolute value squared of the resulting spectra, thus yielding a vector for each time step and a two-dimensional array, like an image, cf. [18].

Reproducible code and more detailed information of our computational experiments can be found in the online repository [25].

7.1 Data and Objective

The publicly available GoodSounds data set [42] contains recordings of single notes and scales played by several single instruments. To gain equally balanced input classes, we restrict the classification problem to six instruments: clarinet, flute, trumpet, violin, alto saxophone and cello. Note that the recordings are monophonic, so that each recording yields one spectrogram that we aim to correctly assign to one of the six instruments.

After removing the silence [3,38], segments from the raw audio files are transformed into log-mel spectrograms [36], so that we obtain images of time–frequency representations with size 100×100. One example spectrogram for each class of instruments is depicted in Fig. 7.

7.2 Convolutional Neural Network Learning

We implemented a fully convolutional neural network $f_\theta : \mathbb{R}^r \to [0,1]^s$, cf. [33], where $r = 100 \times 100$ and $s = 6$, in Python 3.6 using Keras 2.2.4 framework [14] and trained it on the Nvidia GTX 1080 Ti GPU. The data are split into $1,40,722$ training, $36{,}000$ validation and $36{,}000$ independent test samples. We heuristically choose $d = 16$ output features arising directly from the particular output class. The transformations $T_1, \ldots, T_{16}$, with $T_j : \mathbb{R}^6 \to \mathbb{R}$ for $j = 1, \ldots, 16$, are then given by the inner product of the output/target and the feature vectors. Among others, the features are chosen from the enhanced scheme of taxonomy [53] and from the table of frequencies, harmonics and under tones [57]. We use the proposed augmented target loss function L_p (5.2), where L_1 corresponds to the categorical cross-entropy

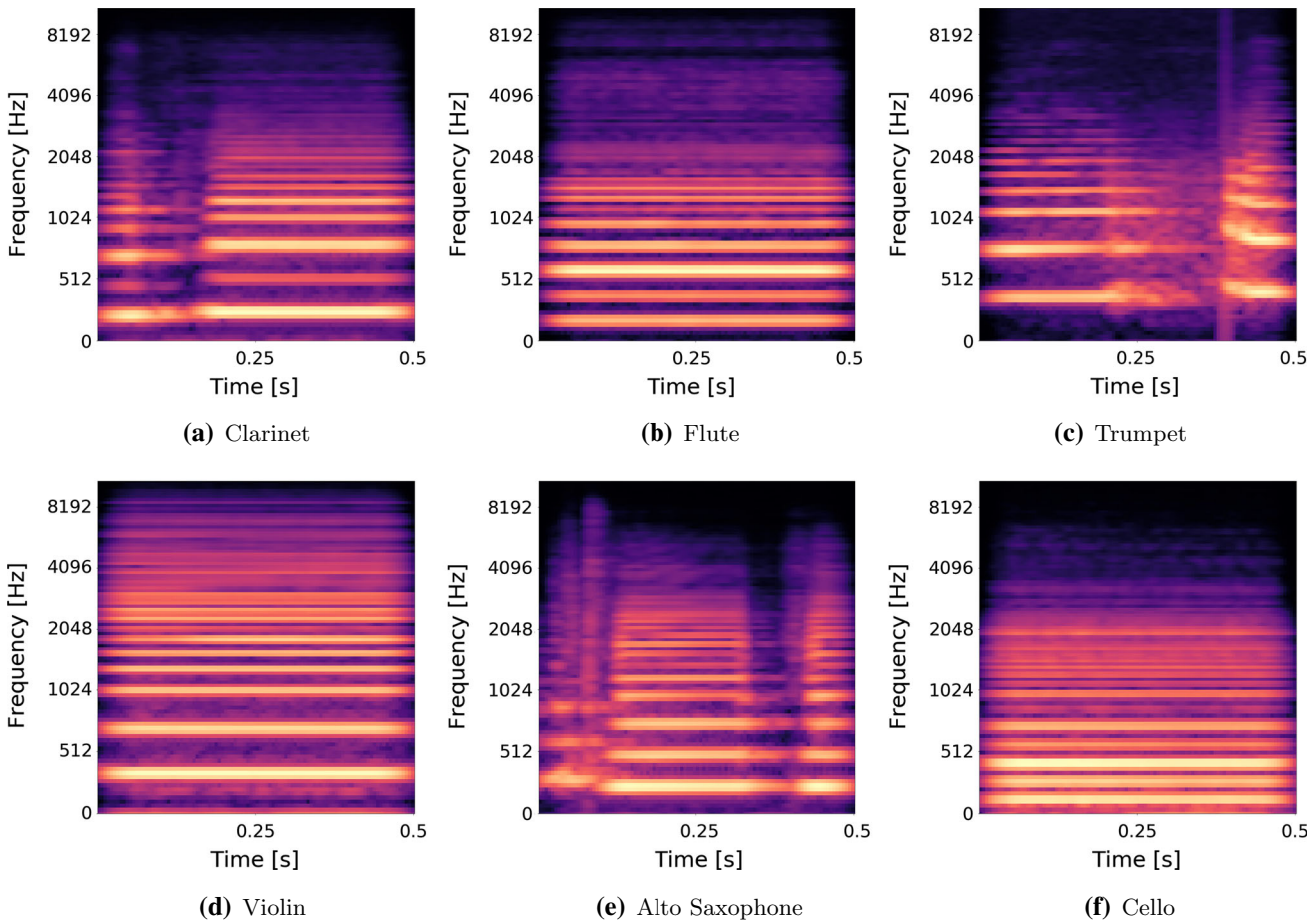


Fig. 7 Log-mel spectrograms of the six different instruments. Intensities range from 0 (black) to 1 (yellow) (Color figure online)

Table 3 Classification results with different parameter choices

α	β	p	Training	Test data
0	0	–	0.5541	0.5716
0.01	0	I_{16}	0.5650	0.5683
0.01	0	$p_{\lambda_{6,16}}$	0.7722	**0.7657**
0	0.05	–	0.9771	0.9729
0.01	0.05	I_{16}	0.9849	0.9802
0.01	0.05	$p_{\lambda_{6,16}}$	0.9857	**0.9833**

The standard inbuilt Tikhonov regularization (ℓ_2-norm of θ) is weighted by β. For $\alpha > 0$, the feature transformations $\{T_j\}_{j=1}^{16}$ are used in the loss function, either directly or weighted by a random projection $p_{\lambda_{6,16}}$. The accuracy of the model is measured by the number of correctly classified samples divided by the number of all samples
Bold values correspond to the highest classification accuracy

loss [55] and L_2 to the mean squared error as in (5.3). We consider here two choices of p: the identity I_{16} and random projectors $p \sim \lambda_{6,16}$ in $\mathcal{G}_{6,16}$.

The deep learning model is sensitive to various hyper-parameters, including α and p, in addition to conventional parameters, such as the number of convolutional kernels, learning rate and the parameter β for Tikhonov regularization. To find the best choices in a fair trial, we utilize a random hyper-parameter search approach, where we train 60 models and select the three best ones for a more precise search over different α in the augmented target loss function and β for Tikhonov regularization. This results in 212 models that are evaluated on the training and validation set. Finally, we select the best model based on the accuracy of the validation set and evaluate it on the independent test set. For comparison, we also evaluate this model with no Tikhonov regularization, i.e., $\beta = 0$; see Table 3.

7.3 Results

Table 3 shows that no regularization and no features provide the poorest results. It seems that adding features with random projections has a regularizing effect and improves the results significantly. As expected, it is important to include Tikhonov regularization on θ. Further enhancement happens by adding features via the modified augmented target loss function with or without additional weighting from projections. All results are very stable and are generalizing very well from training to the independent test set; see [25] for further details.

Acknowledgements Open access funding provided by University of Vienna. This work was partially funded by the Vienna Science and Technology Fund (WWTF) through project VRG12-009, by WWTF AugUniWien/FA746A0249, by International Mobility of Researchers (CZ.02.2.69/0.0/0.0/16 027/0008371) and by project LO1401. For the research, infrastructure of the SIX Center was used.

Appendix A: Proof of Theorem 2.5

A.1 Proof of (2.14) in Theorem 2.5

For $\{y_i\}_{i=1}^M \subset \mathbb{R}^d$ and $p \in \mathcal{G}_{k,d}$, we define

$$f(p, \{y_i\}_{i=1}^M) := \frac{1}{M} \sum_{i=1}^{M} \frac{d}{k} \|p(y_i)\|^2. \tag{A.1}$$

Given two sets, $\{y_i\}_{i=1}^{M_1}, \{z_j\}_{j=1}^{M_2} \subset \mathbb{R}^d$, suppose that $P \in \mathcal{G}_{k,d}$ is a random matrix, distributed according to a cubature measure of strength at least 2. The covariance is given by

$$\begin{aligned}&\mathrm{Cov}(f(P, \{y_i\}_{i=1}^{M_1}), f(P, \{z_j\}_{j=1}^{M_2})) \\ &\quad = \mathbb{E}[(f(P, \{y_i\}) - \mathbb{E}[f(P, \{y_i\})])(f(P, \{z_i\}) \\ &\qquad - \mathbb{E}[f(P, \{z_i\})]\end{aligned}$$

Using the identity, cf. [2],

$$\frac{d}{k}\mathbb{E}\big[\, \|Py\|^2 \,\big] = \|y\|^2 \tag{A.2}$$

directly yields

$$\begin{aligned}&\mathrm{Cov}(f(P, \{y_i\}_{i=1}^{M_1}), f(P, \{z_j\}_{j=1}^{M_2})) \\ &\quad = \mathbb{E}\left[\left(\frac{1}{M_1}\sum_{i=1}^{M_1} \frac{d}{k}\|P(y_i)\|^2 - \frac{1}{M_1}\sum_{i=1}^{M_1}\|y_i\|^2\right)\right. \\ &\qquad \left.\left(\frac{1}{M_2}\sum_{i=1}^{M_2} \frac{d}{k}\|P(z_i)\|^2 - \frac{1}{M_2}\sum_{i=1}^{M_2}\|z_i\|^2\right)\right].\end{aligned}$$

Following [8, Theorem 2.4, Sect. 3.1], we use that

$$\mathbb{E}\big[\, \|Py\|^2 \, \|Pz\|^2\big] = \frac{1}{q}(\alpha_1 \|y\|^2 \|z\|^2 + \alpha_2 \langle y, z\rangle^2), \quad y, z \in \mathbb{R}^d, \tag{A.3}$$

holds, where $q = (d-1)d(d+2)$, $\alpha_1 = (d+1)k^2 - 2k$ and $\alpha_2 = 2k(d-k)$. This leads to the explicit formula of the population covariance

$$\begin{aligned}&\mathrm{Cov}\left(f\left(P, \{y_i\}_{i=1}^{M_1}\right), f\left(P, \{z_j\}_{j=1}^{M_2}\right)\right) \\ &\quad = \frac{a_{k,d}}{M_1 M_2}\sum_{i=1}^{M_1}\sum_{j=1}^{M_2}\langle y_i, z_j\rangle^2\end{aligned}$$

$$- \frac{a_{k,d}}{d} \left(\frac{1}{M_1} \sum_{i=1}^{M_1} \|y_i\|^2 \right) \left(\frac{1}{M_2} \sum_{j=1}^{M_2} \|z_j\|^2 \right), \tag{A.4}$$

with $a_{k,d} = \frac{2d(d-k)}{k(d-1)(d+2)}$.
For $y := \{y_i\}_{i=1}^M \subset \mathbb{R}^d \backslash \{0\}$, we set $\hat{y}_i := \frac{y_i}{\|y_i\|}$, for $i = 1, \ldots, M$. The identity (A.4) enables us to compute the population correlation

$$\mathrm{Corr}(f(P, y), f(P, \hat{y})) = \frac{\mathrm{Cov}(f(P, y), f(P, \hat{y}))}{\sqrt{\mathrm{Var}(f(P, y))}\sqrt{\mathrm{Var}(f(P, \hat{y}))}} \tag{A.5}$$

by the explicit formulas

$$\mathrm{Cov}\left[f(P, y), f(P, \hat{y})\right] = \frac{a_{k,d}}{M^2} \sum_{i,j=1}^{M} \langle y_i, \hat{y}_j \rangle^2 - \frac{a_{k,d}}{d} \cdot \frac{1}{M} \sum_{i=1}^{M} \|y_i\|^2$$

$$\mathrm{Cov}\left[f(P, y), f(P, y)\right] = \mathrm{Var}[f(P, y)] = \frac{a_{k,d}}{M^2} \sum_{i,j=1}^{M} \langle y_i, y_j \rangle^2 - \frac{a_{k,d}}{d} \left(\frac{1}{M} \sum_{i=1}^{M} \|y_i\|^2 \right)^2$$

$$\mathrm{Cov}\left[f(P, \hat{y}), f(P, \hat{y})\right] = \mathrm{Var}[f(P, \hat{y})] = \frac{a_{k,d}}{M^2} \sum_{i,j=1}^{M} \langle \hat{y}_i, \hat{y}_j \rangle^2 - \frac{a_{k,d}}{d}.$$

Since the variance is always nonnegative and $\frac{a_{k,d}}{d} > 0$, the denumerator of $\mathrm{Corr}(f(P, y), f(P, \hat{y}))$ in (A.5) satisfies

$$\begin{aligned}
&\sqrt{\mathrm{Var}(f(P, y))}\sqrt{\mathrm{Var}(f(P, \hat{y}))} \\
&\leq \sqrt{\left(\frac{a_{k,d}}{M^2} \sum_{i,j=1}^{M} \langle y_i, y_j \rangle^2 \right) \left(\frac{a_{k,d}}{M^2} \sum_{i,j=1}^{M} \langle \hat{y}_i, \hat{y}_j \rangle^2 \right)} \\
&\leq \frac{a_{k,d}}{M^2} \sqrt{\left(\sum_{i,j=1}^{M} \langle y_i, y_j \rangle^2 \right) \left(\frac{1}{\min_i (\|y_i\|)^4} \sum_{i,j=1}^{M} \langle y_i, y_j \rangle^2 \right)} \\
&\leq \frac{1}{\min_i (\|y_i\|)^2} \frac{a_{k,d}}{M^2} \sum_{i,j=1}^{M} \langle y_i, y_j \rangle^2.
\end{aligned}$$

The enumerator of $\mathrm{Corr}(f(P, y), f(P, \hat{y}))$ in (A.5) is estimated by

$$\mathrm{Cov}(f(P, y), f(P, \hat{y})) \geq \frac{a_{k,d}}{\max_i (\|y_i\|)^2} \frac{1}{M^2} \sum_{i,j=1}^{M} \langle y_i, y_j \rangle^2 - \frac{a_{k,d}}{d} \max_i (\|y_i\|)^2.$$

For $d \geq M$, a short calculation yields $\mathrm{Cov}(f(P, y), f(P, \hat{y})) \geq 0$, so that we obtain

$$\mathrm{Corr}(f(P, y), f(P, \hat{y})) \geq \frac{\min_i (\|y_i\|)^2}{\max_i (\|y_i\|)^2} - \frac{\min_i (\|y_i\|)^2 \max_i (\|y_i\|)^2}{\frac{d}{M^2} \sum_{i,j=1}^{M} \langle y_i, y_j \rangle^2}.$$

The lower bound $\sum_{i,j=1}^{M} \langle y_i, y_j \rangle^2 \geq M \min_i (\|y_i\|)^4$ yields

$$\mathrm{Corr}(f(P, y), f(P, \hat{y})) \geq \frac{\min_i (\|y_i\|)^2}{\max_i (\|y_i\|)^2} - \frac{M}{d} \cdot \frac{\max_i (\|y_i\|)^2}{\min_i (\|y_i\|)^2}.$$

Since the correlation is scaling invariant, the choice $y = \{x_i - x_j : 1 \leq i < j \leq m\}$ with $M = \frac{m(m-1)}{2}$ implies (2.14) in Theorem 2.5. Incorporating the correct scaling yields the following corollary:

Corollary A.1 *For a given data set $x = \{x_i\}_{i=1}^m$ and for random $P \in \mathcal{G}_{k,d}$ the (co)variances of* $\mathrm{tvar}(Px)$ (2.4) *and* $\mathcal{M}(P, x)$ (2.8) *are given by*

$$\mathrm{Cov}(\mathcal{M}(P, x), \mathrm{tvar}(Px)) = \frac{k}{2d} \left(\frac{a_{k,d}}{M^2} \sum_{i<j} \sum_{l<r} \left\langle x_i - x_j, \frac{x_l - x_r}{\|x_l - x_r\|} \right\rangle^2 - \frac{a_{k,d}}{d} \cdot \frac{1}{M} \sum_{i<j} \|x_i - x_j\|^2 \right),$$

$$\mathrm{Var}(\mathrm{tvar}(Px)) = \frac{k^2}{4d^2} \left(\frac{a_{k,d}}{M^2} \sum_{i<j} \sum_{l<r} \langle x_i - x_j, x_l - x_r \rangle^2 - \frac{a_{k,d}}{d} \left(\frac{1}{M} \sum_{i<j} \|x_i - x_j\|^2 \right)^2 \right),$$

$$\mathrm{Var}(\mathcal{M}(P, x)) = \frac{a_{k,d}}{M^2} \sum_{i<j} \sum_{l<r} \left\langle \frac{x_i - x_j}{\|x_i - x_j\|}, \frac{x_l - x_r}{\|x_l - x_r\|} \right\rangle^2 - \frac{a_{k,d}}{d},$$

where $M = \frac{m(m-1)}{2}$ *and* $a_{k,d} = \frac{2d(d-k)}{k(d-1)(d+2)}$.

A.2 Proof of the Second Part of Theorem 2.5

For fixed parameters $\mu > 0, \sigma^2 > 0$, that do not depend on d, let $Y_1 \in \mathbb{R}^d$ be a random vector, whose squared entries are independent, identically distributed with mean $\mathbb{E}Y_{1,l}^2 = \mu$ and variance $\mathrm{Var}(Y_{1,l}^2) = \sigma^2$, for $l = 1, \ldots, d$. We immediately observe

$$\mathbb{E}\left(\frac{\|Y_1\|^2}{\sqrt{d}} \right) = \sqrt{d}\mu, \qquad \mathrm{Var}\left(\frac{\|Y_1\|^2}{\sqrt{d}} \right) = \sigma^2.$$

For any $c > 0$, Chebyshev's inequality yields

$$\mathbb{P}\left(\left| \frac{\|Y_1\|^2}{\sqrt{d}} - \sqrt{d}\mu \right| \geq c\sigma \right) \leq \frac{1}{c^2}.$$

Suppose that $Y_2, \ldots, Y_M$ are copies of Y_1, not necessarily independent. Then, the union bound

$$\mathbb{P}\Big(\Big|\frac{\|Y_i\|^2}{\sqrt{d}} - \sqrt{d}\mu\Big| \geq c\sigma, \text{ for some } i = 1, \ldots, M\Big) \leq \frac{M}{c^2}$$

implies that

$$\sqrt{d}\mu - c\sigma \leq \frac{\min_i(\|Y_i\|)^2}{\sqrt{d}} \leq \frac{\max_i(\|Y_i\|)^2}{\sqrt{d}} \leq \sqrt{d}\mu + c\sigma$$

holds with probability at least $1 - \frac{M}{c^2}$. Provided that $\sqrt{d}\mu \neq c\sigma$ and $0 < \sqrt{d}\mu - c\sigma$, we deduce

$$\frac{\sqrt{d}\mu - c\sigma}{\sqrt{d}\mu + c\sigma} \leq \frac{\min_i(\|Y_i\|)^2}{\max_i(\|Y_i\|)^2} \leq \frac{\sqrt{d}\mu + c\sigma}{\sqrt{d}\mu - c\sigma}.$$

We can choose $c = \frac{\mu}{\sigma}\sqrt[4]{d}$, since $0 < c \leq \frac{\sqrt[4]{d}\mu}{\sigma} \leq \frac{\sqrt{d}\mu}{\sigma}$. That directly yields

$$\frac{1 - \frac{1}{\sqrt[4]{d}}}{1 + \frac{1}{\sqrt[4]{d}}} \leq \frac{\min_i(\|Y_i\|)^2}{\max_i(\|Y_i\|)^2} \leq \frac{1 + \frac{1}{\sqrt[4]{d}}}{1 - \frac{1}{\sqrt[4]{d}}}$$

and holds with probability at least $1 - \frac{\mu^2 M}{\sigma^2\sqrt{d}}$.

It follows directly that $\frac{\min_i(\|Y_i\|)^2}{\max_i(\|Y_i\|)^2}$ converges toward 1 in probability for $d \to \infty$,

The choice $\{Y_1, \ldots, Y_M\} = \{X_i - X_j : 1 \leq i < j \leq m\}$ implies the second part of Theorem 2.5.

A.3 Calculations for Population Covariances

We notice that $\|p(x_i - x_j)\|^2 = \text{trace}(px_ix_i^\top - px_jx_j^\top)$ is a polynomial of degree 1 in p. Hence, $\text{tvar}(px)$ in (2.4) is also a polynomial of degree 1 in p. If $\{p_l\}_{l=1}^n$ is a 1-design, then the sample mean of $\{\text{tvar}(p_1x), \ldots, \text{tvar}(p_nx)\}$ satisfies

$$\frac{1}{n}\sum_{l=1}^{n}\text{tvar}(p_lx) = \mathbb{E}\,\text{tvar}(Px),$$

which is the population mean of $\text{tvar}(Px)$, with $P \sim \lambda_{k,d}$. Similarly, the term $\|p(x_i - x_j)\|^4$ is a polynomial of degree 2 in p, so that $(\mathcal{M}(p, x))^2$ in (2.8) is a polynomial of degree 2 in p. If $\{p_l\}_{l=1}^n$ is a 2-design, then we derive

$$\sum_{l=1}^{n}(\mathcal{M}(p_l, x))^2 - \left(\sum_{j=1}^{n}\mathcal{M}(p_l, x)\right)^2$$
$$= \mathbb{E}(\mathcal{M}(P, x))^2 - \mathbb{E}\left(\sum_{j=1}^{n}\mathcal{M}(P, x)\right)^2,$$

with $P \sim \lambda_{k,d}$. In other words, the sample variance of $\{\mathcal{M}(p_1, x), \ldots, \mathcal{M}(p_n, x)\}$ coincides with the population variance $\text{Var}(\mathcal{M}(P, x))$. Analogously, we deduce that the sample covariance of (3.5) coincides with the population covariance $\text{Cov}(\mathcal{M}(P, x), \text{tvar}(Px))$ with $P \sim \lambda_{k,d}$.

References

1. Achlioptas, D.: Database-friendly random projections: Johnson–Lindenstrauss with binary coins. J. Comput. Syst. Sci. **66**(4), 671–687 (2003)
2. Bachoc, C., Ehler, M.: Tight p-fusion frames. Appl. Comput. Harmon. Anal. **35**(1), 1–15 (2013)
3. Bagwell, C.: SoX: Sound eXchange the Swiss army knife of sound processing. https://launchpad.net/ubuntu/+source/sox/14.4.1-5. Accessed 31 Oct 2018
4. Ball, K.: An elementary introduction to modern convex geometry. Flavors Geom. **31**, 1–58 (1997)
5. Baraniuk, R.G., Wakin, M.B.: Random projections of smooth manifolds. Found. Comput. Math. **9**, 941–944 (2006)
6. Baraniuk, R., Davenport, M., DeVore, R., Wakin, M.: A simple proof of the restricted isometry property for random matrices. Constr. Approx. **28**(3), 253–263 (2008)
7. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: applications to image and text data. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA), KDD '01, ACM, pp. 245–250 (2001)
8. Bodman, B., Ehler, M., Gräf, M.: From low to high-dimensional moments without magic. J. Theor. Probab. **31**(4), 2167–2193 (2017)
9. Breger, A., Ehler, M., Bogunovic, H., Waldstein, S.M., Philip, A., Schmidt-Erfurth, U., Gerendas, B.S.: Supervised learning and dimension reduction techniques for quantification of retinal fluid in optical coherence tomography images, Eye, Springer Nature (2017)
10. Breger, A., Ehler, M., Gräf, M.: Quasi Monte Carlo Integration and Kernel-based Function Approximation on Grassmannians, Frames and Other Bases in Abstract and Function Spaces, Applied and Numerical Harmonic Analysis Series (ANHA). Springer, Birkhauser (2017)
11. Breger, A., Ehler, M., Gräf, M.: Points on manifolds with asymptotically optimal covering radius. J. Complex. **48**, 1–14 (2018)
12. Candès, E.J., Tao, T.: Decoding by linear programming. IEEE Trans. Inf. Theory **51**(12), 4203–4215 (2005)
13. Caruana, R.: Multitask learning. Mach. Learn. **28**(1), 41–75 (1997)
14. Chollet, F., et al.: Keras (2015) https://keras.io
15. Dasgupta, S., Gupta, A.: An elementary proof of a theorem of Johnson and Lindenstrauss. Random Struct. Algorithms **22**(1), 60–65 (2003)
16. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd International Conference on Machine Learning (New York, NY, USA), ICML '06, ACM, pp. 233–240 (2006)
17. de la Harpe, P., Pache, C.: Cubature formulas, geometrical designs, reproducing kernels, and Markov operators, Infinite groups: geometric, combinatorial and dynamical aspects (Basel), vol. 248, Birkhäuser, pp. 219–267 (2005)
18. Dörfler, M., Bammer, R., Grill, T.: Inside the spectrogram: convolutional neural networks in audio processing. In: IEEE International Conference on Sampling Theory and Applications (SampTA), pp. 152–155 (2017)

19. Dörfler, M., Grill, T., Bammer, R., Flexer, A.: Basic filters for convolutional neural networks applied to music: training or design. Neural Comput. Appl. 1–14 (2018)
20. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems (USA), NIPS'16, Curran Associates Inc., pp. 658–666 (2016)
21. Etayo, U., Marzo, J., Ortega-Cerdà, J.: Asymptotically optimal designs on compact algebraic manifolds. J. Monatsh. Math. **186**(2), 235–248 (2018)
22. Ganaye, P.-A., Sdika, M., Benoit-Cattin, H.: Semi-supervised learning for segmentation under semantic constraint. In: 21st International Conference, Granada, Spain, Sept 16–20, 2018, Proceedings, Part III, pp. 595–602 (2018)
23. Gerendas, B.S., Hu, X., Kaider, A., Montuoro, A., Sadeghipour, A., Waldstein, S.M., Schmidt-Erfurth, U.: Oct biomarkers predictive for visual acuity in patients with diabetic macular edema. Investig. Ophthalmol. Vis. Sci. **58**(8), 2026–2026 (2017)
24. Golub, G.H., Van Loan, C.F.: Matrix Computations, Johns Hopkins Studies in the Mathematical Sciences. The Johns Hopkins University Press (1996)
25. Harar, P.: Orthovar (2018) https://gitlab.com/hararticles/orthovar
26. Heckel, R., Tschannen, M., Bölcskei, H.: Dimensionality-reduced subspace clustering. Inf. Inference: J. IMA **6**, 246–283 (2017)
27. Hedge, C., Sankaranarayanan, A.C., Yin, W., Baraniuk, R.G.: Numax: a convex approach for learning near-isometric linear embeddings. IEEE Trans. Signal Process. **83**, 6109–6121 (2015)
28. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016)
29. Karr, J.R., Martin, T.E.: Random numbers and principal components: further searches for the unicorn, Tech. report, United States Forest Service General Technical Report (1981)
30. Krahmer, F., Ward, R.: New and improved Johnson Lindenstrauss embeddings via the restricted isometry property. SIAM J. Math. Anal. **43**(3), 1269–1281 (2011)
31. Ledig, C., Theis, L., Huszar, F., Caballero, J., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., Shi W.: Photo-realistic single image super-resolution using a generative adversarial network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 105–114 (2017)
32. Linial, N., London, E., Rabinovich, Y.: The geometry of graphs and some of its algorithmic applications. Combinatorica **15**(2), 215–245 (1995)
33. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
34. Mahoney, M.W.: Randomized algorithms for matrices and data. Found. Trends® Mach. Learn. **3**(2), 123–224 (2011)
35. Matousek, J.: On variants of the Johnson–Lindenstrauss lemma. Random Struct. Algorithms **33**(2), 142–156 (2008)
36. McFee, B., et al.: Librosa: 0.6.2 (2018) https://doi.org/10.5281/zenodo.1342708
37. Moore, B.: Principal component analysis in linear systems: controllability, observability, and model reduction. IEEE Trans. Autom. Control **26**, 17–32 (1981)
38. Navarrete, J.: The sox of silence (2009) https://digitalcardboard.com/blog/2009/08/25/the-sox-of-silence
39. Neumayer, S., Nimmer, M., Setzer, S., Steidl, G.: On the robust PCA and Weiszfeld's algorithm.Appl. Math. Optim. 1–32 (2019)
40. Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Guerrero, R., Cook, S.A., de Marvao, A., Dawes, T., O'Regan, D., Kainz, B., Glocker, B., Rueckert, D.: Anatomically constrained neural networks (ACNN): application to cardiac image enhancement and segmentation. IEEE Trans. Med. Imaging **37**, 384–395 (2018)
41. Pabst, G.: Parameters for Compartment-Free Pharmacokinetics-Standardisation of Study Design, Data Analysis and Reporting, ch. 5. Area Under the Concentration-Time Curve, pp. 65–80, Shaker Verlag (1999)
42. Picas, O.R., Rodriguez, H.P., Dabiri, D., Tokuda, H., Hariya, W., Oishi, K., Serra, X.: A real-time system for measuring sound goodness in instrumental sounds. In: Audio Engineering Society Convention 138, Audio Engineering Society (2015)
43. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation (2015) arXiv:1505.04597
44. Seymour, P., Zaslavsky, T.: Averaging sets: a generalization of mean values and spherical designs. Adv. Math. **52**, 213–240 (1984)
45. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014) arXiv:1409.1556
46. Stauffer, D.F., Garton, E.O., Steinhorst, R.K.: Ecology: a comparison of principal components from real and random data. Ecology **66**(6), 1693–1698 (1985)
47. Stewart, R., Ermon, S.: Label-free supervision of neural networks with physics and domain knowledge. In: AAAI (2017)
48. Thanei, G.-A., Heinze, C., Meinshausen, N.: Random Projections for Large-scale Regression, pp. 51–68. Springer, Cham (2017)
49. Udell, M.: Generalized low rank models, Ph.D. thesis, Stanford University (2015)
50. Varnousfaderani, E.S., Wu, J., Vogl, W.-D., Philip, A.-M., Montuoro, A., Leitner, R., Simader, C., Waldstein, S.M., Gerendas, B.S., Schmidt-Erfurth, U.: A novel benchmark model for intelligent annotation of spectral-domain optical coherence tomography scans using the example of cyst annotation. Comput. Methods Programs Biomed. **130**, 93–105 (2016)
51. Veraart, Jelle, Novikov, Dmitry S., Christiaens, Daan, Ades-aron, Benjamin, Sijbers, Jan, Fieremans, Els: Denoising of diffusion MRI using random matrix theory. NeuroImage **142**, 394–406 (2016)
52. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. In: Eldar, Y., Kutyniok, G. (eds.) Compressed Sensing, Theory and Applications, pp. 210–268. Cambridge University Press, Cambridge (2012)
53. von Hornbostel, E.M., Sachs, C.: Classification of musical instruments: translated from the original german by anthony baines and klaus p. wachsmann. Galpin Soc. J. 3–29 (1961)
54. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 2 (Cambridge, MA, USA), NIPS'14, MIT Press, pp. 3320–3328 (2014)
55. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: Advances in Neural Information Processing Systems (NIPS) **31** (2018)
56. Zhang, L., Lukac, R., Wu, X., Zhang, D.: Pca-based spatially adaptive denoising of cfa images for single-sensor digital cameras. IEEE Trans. Image Process. **18**(4), 797–812 (2009)
57. ZyTrax Inc.: Frequency ranges (2018) http://www.zytrax.com/tech/audio/audio.html

A. Breger received her MSc degree in Mathematics from the University of Vienna in 2015, followed by a research position in collaboration with the Medical University of Vienna. She focuses on interdisciplinary research with her PhD studies on 'Dimension reduction with orthogonal projections and applications on medical images with learning tasks'. In 2019 she received a grant to work at the Laboratory of Mathematics in Imaging (LMI), Brigham and Women's Hospital, Harvard Medical School, Boston (MA).

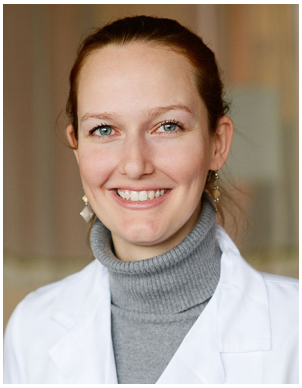
J. I. Orlando obtained his Ph.D. in computational and industrial mathematics from Universidad Nacional del Centro de la Provincia de Buenos Aires (UNICEN), Tandil, Argentina. In 2018 and 2019, he has worked as a Postdoctoral Researcher at the Department of Ophthalmology of the Medical University of Vienna (Austria). He is currently a permanent researcher from the National Scientific and Technical Research Council (CONICET, Argentina), working at the research group Yatiris from Pladema Institute (UNICEN). His primary research interests include machine learning and computer vision applied to computer-assisted ophthalmology and medicine in general.

P. Harar obtained a M.Sc. in System Engineering and Informatics from Brno University of Technology in 2015 before pursuing a Ph.D. at the Brain Diseases Analysis Laboratory on "Audio Classification with Deep Learning on Limited Data Sets". He has been working at the Numerical Harmonic Analysis Group (University of Vienna) since 2018. His research focuses on medical audio analysis using deep learning methods and their development.

M. Dörfler obtained her PhD in Mathematics from the University of Vienna and is a research assistant at the Mathematics Department of the University of Vienna. She is heading her own research group and is interested in the interplay between aspects of harmonic analysis and the design of learning algorithms based on convolutional neural networks. She has worked on audio signal processing and, being a musician herself, has been particularly interested in developing representations adapted to music applications.

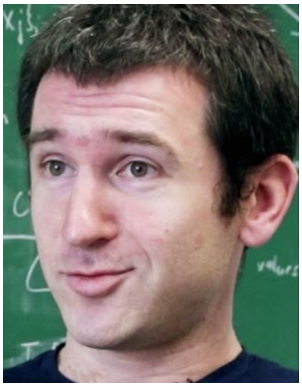
S. Klimscha is a team member of the OPTIMA lab and a resident at the Department of Ophthalmology, Medical University of Vienna. She obtained her MD degree at the Medical University of Vienna, Austria in July 2015, joined the OPTIMA team in January 2016 and is enrolled in the PhD Program for Medical Imaging.

C. Grechenig obtained his MD degree at the Medical University of Graz, Austria in January 2017, and joined the Christian Doppler Laboratory for Ophthalmic Image Analysis (OPTIMA) in March 2018 and is enrolled in the PhD Program for Medical Imaging at the Medical University of Vienna. As a clinician he currently works as a retina resident at the Department of Ophthalmology and Optometry, Medical University of Vienna.



B. S. Gerendas is the Managing Director of the Vienna Reading Center at the Department of Ophthalmology, Medical University of Vienna since 2017. She completed Medical School in 2009 at the University of Heidelberg, Germany, where she in parallel obtained a Master of Science degree in Healthcare Management. Besides her PhD in Medical Physics, she is a board certified ophthalmologist/retina specialist and finished her habilitation (Assoc. Prof.) in 2018. Her primary research interests are in the field of retinal and choroidal imaging and (automated) image analysis as well as translational ophthalmic research in retinal diseases.

U. Schmidt-Erfurth is Chair of the Department of Ophthalmology at the Medical University of Vienna, Austria. Her areas of expertise include surgical and medical retina and the development of innovative diagnostic techniques and treatment strategies for chorioretinal diseases. She is the Head of the Vienna Reading Center and the founder and leader of the OPTIMA project which introduces means of artificial intelligence into retinal imaging. She is the author of over 420 original articles. She has received numerous grants/awards and is a member of the American Academy of Ophthalmology, EURETINA, the European and the Austrian Academy of Sciences.



M. Ehler After his PhD in mathematics in 2007, he received a 3 year PostDoctoral Fellowship at the National Institutes of Health in Bethesda/Maryland. His research efforts are guided by bridging gaps between theoretical and numerical mathematics and medical applications. Since 2013 he is the head of a Vienna Research Group for Young Investigators on computational harmonic analysis of high-dimensional biomedical data. At the end of 2017 he became Assoc. Professor at the Faculty of Mathematics at the University of Vienna.