SqueezeFit: Label-aware dimensionality reduction by semidefinite programming

Culver McWhirter* Dustin G. Mixon* Soledad Villar†

Abstract

Given labeled points in a high-dimensional vector space, we seek a low-dimensional subspace such that projecting onto this subspace maintains some prescribed distance between points of differing labels. Intended applications include compressive classification. Taking inspiration from large margin nearest neighbor classification, this paper introduces a semidefinite relaxation of this problem. Unlike its predecessors, this relaxation is amenable to theoretical analysis, allowing us to provably recover a planted projection operator from the data.

1 Introduction

The last decade of sampling theory has transformed the way we reconstruct signals from measurements. For example, the now-established theory of compressed sensing allows one to reconstruct a signal from a number of random linear measurements that is proportional to the complexity of that signal [12, 9, 14], potentially speeding up MRI scans by a factor of five [27]. This theory has since transferred to the setting of nonlinear measurements in the context of phase retrieval [10, 7], leading to new algorithms for coherent diffractive imaging [33]. Today, we witness major technological advances in machine learning, where neural networks have recently achieved unprecedented performance in image classification and elsewhere [22, 34]. This motivates another fundamental problem for sampling theory:

How many samples are necessary to enable signal classification?

For instance, why waste time collecting enough samples to completely reconstruct a given signal if you only need to detect whether the signal contains an anomaly?

This different approach to sampling is known as **compressive classification**. While the idea has been around since 2007, to date, only three works provide theory to derive sampling rates for compressive classification. First, [11] considered the case where each class is a low-dimensional manifold. Much later, [30] compressively classified mixtures of Gaussians of low-rank covariance, and then [3] derived sampling rates for random projection to maintain separation between full-dimensional ellipsoids. Overall, these works assumed that the classes follow a specific model (be it manifolds, Gaussians or ellipsoids), and then derived conditions under which a good projection exists. The present work takes a dual approach: We assume that compressive classification is possible, meaning there exists a planted low-rank projection that facilitates classification, and the task is to derive conditions on the classes for which finding that projection is feasible:

*The Ohio State University, Columbus, OH 43210
†New York University, New York, NY 10003

Problem 1 (projection factor recovery)**.** Let Π denote orthogonal projection onto some unknown subspace $T \subseteq \mathbb{R}^d$ of some unknown dimension. What conditions on $f\colon T \to [k] := \{1, \ldots, k\}$ and $\mathcal{X} \subseteq \mathbb{R}^d$ enable exact or approximate recovery of Π from data of the form $\{(x, f(\Pi x))\}_{x \in \mathcal{X}}$?

In words, we assume the classification function factors through some unknown orthogonal projection operator Π, and the objective is to reconstruct Π. Once we find Π of rank r, then we may write $\Pi = A^\top A$ for some $r \times d$ sensing matrix A, and then Ax determines the classification $f(\Pi x)$ of x despite using only $r \ll d$ samples. Here and throughout, we consider a sequence of data $\mathcal{D} = \{(x_i, y_i)\}_{i \in \mathcal{I}}$ in $\mathbb{R}^d \times [k]$ and denote $\mathcal{Z}(\mathcal{D}) := \{x_i - x_j : i, j \in \mathcal{I}, y_i \neq y_j\}$. The following program finds the best orthogonal projection for our purposes:

$$\text{minimize} \quad \operatorname{rank} \Pi \quad \text{subject to} \quad \|\Pi z\| \geq \Delta \;\; \forall z \in \mathcal{Z}(\mathcal{D}), \quad \Pi^\top = \Pi, \quad \Pi^2 = \Pi \tag{1}$$

Here, Π is the decision variable, whereas $\Delta > 0$ is a parameter that prescribes a desired minimum distance between projected points Πx_i and Πx_j with differing labels. This parameter reflects a fundamental tension in compressive classification: We want Δ to be large so as to enable classification, but we also want $\operatorname{rank} \Pi$ to be small so that this classification is compressive. Since it is not clear how to tractably implement (1), we consider a convex relaxation:

$$\text{minimize} \quad \operatorname{tr} M \quad \text{subject to} \quad z^\top M z \geq \Delta^2 \;\; \forall z \in \mathcal{Z}(\mathcal{D}), \quad 0 \preceq M \preceq I \tag{$\operatorname{sqz}(\mathcal{D}, \Delta)$}$$

We refer to this program as **SqueezeFit**. If $\mathcal{Z}(\mathcal{D})$ is finite, then $\operatorname{sqz}(\mathcal{D}, \Delta)$ is a semidefinite program, otherwise $\operatorname{sqz}(\mathcal{D}, \Delta)$ is a semi-infinite program [16]. In either case, the minimum exists whenever $\operatorname{sqz}(\mathcal{D}, \Delta)$ is feasible by the extreme value theorem. As Figure 1 illustrates, SqueezeFit is well suited for projection factor recovery.

1.1 Relationship to previous work

We note that our formulation of projection factor recovery in Problem 1 can be viewed as an instance of the general problem of learning **multiple-index models (MIMs)**. Specifically, an MIM consists of an unknown matrix $A \in \mathbb{R}^{r \times d}$, a (possibly unknown) link function $f\colon \mathbb{R}^r \to \mathbb{R}$, a random vector $X \in \mathbb{R}^d$, and random noise $Z \in \mathbb{R}$:

$$Y := f(AX) + Z.$$

The learning task is to estimate A or $\operatorname{im}(A^\top)$ from a collection of examples of (X, Y). From this general perspective, projection factor recovery can be thought of as learning an MIM in which f is an unknown classification function. We note that when $r = 1$, the general MIM reduces to a so-called *single-index model*, which has received considerable attention. For example, the case where $f(t) = t^2$ corresponds to phase retrieval [10, 7], and $f(t) = \operatorname{sign}(t)$ corresponds to one-bit sensing [4]. There has been considerably less work in learning MIMs with $r > 1$, see [40] and references therein. However, ideas from semidefinite programming also appear relevant in this context, as [40] leverages the sparse PCA semidefinite program from [38] to learn sparse MIMs.

When formulating SqueezeFit, the authors took inspiration from the **large margin nearest neighbor (LMNN)** algorithm [39], which finds the $d \times d$ matrix $M \succeq 0$ such that $\{(M^{1/2} x_i, y_i)\}_{i \in \mathcal{I}}$ is best conditioned for k-nearest neighbor classification in the Euclidean distance (unlike above, k does not correspond to the number of classes here). To accomplish this, LMNN first identifies for each x_i, the k closest x_j such that $y_j = y_i$; these are called *target neighbors*. Next, x_l is called an *impostor* of x_i under M if x_i has a target neighbor x_j such that $\|M^{1/2}(x_i - x_l)\|^2 \leq \|M^{1/2}(x_i - x_j)\|^2 + 1$. Intuitively, $\{(M^{1/2} x_i, y_i)\}_{i \in \mathcal{I}}$ is well conditioned for k-nearest neighbors if

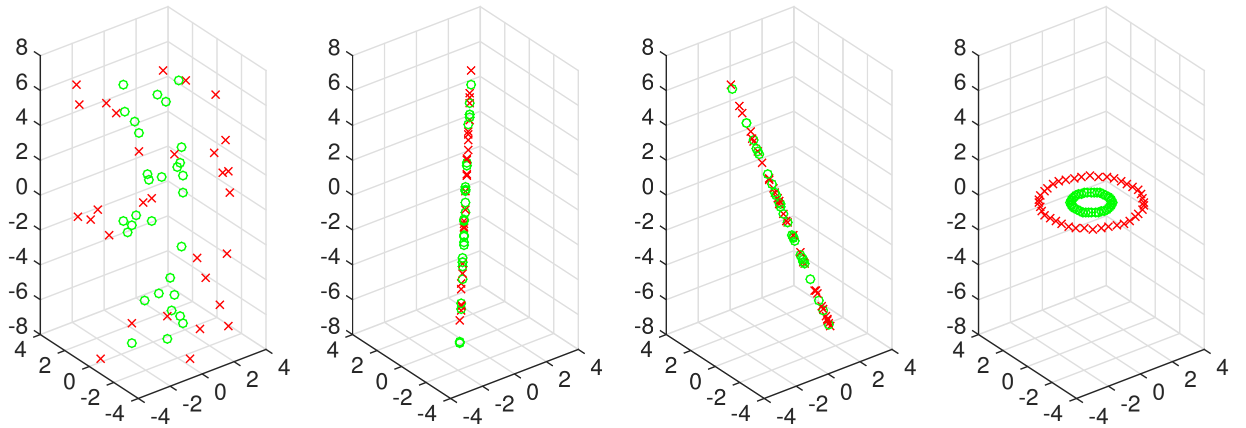

Figure 1: **(far left)** Plot of 60 data points in $\mathbb{R}^3$, half in one class, half in another. These points were drawn according to a random model with an unknown planted projection factor (as in Problem 1). **(middle left)** Principal component analysis (PCA) suggests one-dimensional structure in the data. Projecting onto this subspace (which was identified without regard for the points' classes) results in an undesirable mixture of the classes. **(middle right)** Unlike PCA, linear discriminant analysis (LDA) actually considers which class each point belongs to. Since there are two classes, the result is projection onto a 1-dimensional subspace, obtained by applying the classes' inverse covariance matrix to the difference of class centroids. Unfortunately, the result is again an unhelpful mixture of classes. **(far right)** Unlike PCA and LDA, SqueezeFit finds a low-rank projection that maintains some amount of distance between points from different classes. The resulting projection is a close approximation to the planted projection factor. See Section 2 for theoretical guarantees that help explain this behavior.

the target neighbors are all close to each other, and the number of impostors is small. To this end, LMNN uses a semidefinite program to find the $M \succeq 0$ that simultaneously optimizes these conflicting objectives.

While LMNN has proven to be an effective tool for metric learning, there is currently a dearth of theory to explain its performance. By contrast, our formulation of SqueezeFit is particularly amenable to theoretical analysis, which we credit to two features: First, we do not require a pre-processing step to define target neighbors, thereby isolating how our algorithm depends on the data. In exchange for this lack of pre-processing, we accept the hyperparameter Δ to provide some notion of "impostor." Second, SqueezeFit includes the identity constraint $M \preceq I$, which proves particularly valuable to the theory. For example, the identity constraint plays a key role in the proof that, if M is SqueezeFit-optimal for $\{(x_i, y_i)\}_{i\in\mathcal{I}}$, then the only SqueezeFit-optimal operator for $\{(M^{1/2}x_i, y_i)\}_{i\in\mathcal{I}}$ is orthogonal projection onto $\operatorname{span}\{M^{1/2}x_i\}_{i\in\mathcal{I}}$ (see Theorem 13). In words, *you don't need to squeeze your data more than once.*

Finally, we note that SqueezeFit bears some resemblance to a dimensionality reduction method known as **nuclear norm minimization with max-norm constraints (NuMax)** [18]. This method makes use of inter-class and intra-class secant sets:

$$\mathcal{S}_1(\mathcal{D}) := \{\tfrac{x_i - x_j}{\|x_i - x_j\|} : x_i \neq x_j, y_i \neq y_j\}, \qquad \mathcal{S}_2(\mathcal{D}) := \{\tfrac{x_i - x_j}{\|x_i - x_j\|} : x_i \neq x_j, y_i = y_j\}.$$

For compressive classification, we intuitively want a compressive operator that grows members of

$\mathcal{S}_1(\mathcal{D})$ while simultaneously shrinking members of $\mathcal{S}_2(\mathcal{D})$. This suggests the following program:

$$\text{minimize} \quad \operatorname{tr} M \quad \text{subject to} \quad \begin{array}{l} v^\top M v \geq 1-\delta \;\; \forall v \in \mathcal{S}_1(\mathcal{D}), \\ u^\top M u \leq 1+\delta \;\; \forall u \in \mathcal{S}_2(\mathcal{D}), \quad M \succeq 0 \end{array} \tag{NuMax($\mathcal{D},\delta$)}$$

Perhaps surprisingly, NuMax does not perform well in the context of projection factor recovery. For example, running NuMax on the example in Figure 1 produces a matrix M that is approximately equal to a multiple of the identity. The reason for this behavior stems from the constraints corresponding to the inter-class secants $\mathcal{S}_1(\mathcal{D})$, which we illustrate in the following example:

Example 2. Select a unit vector $w \in \mathbb{R}^n$, and suppose the (infinite) dataset $\mathcal{D} = \{(x_i, y_i)\}_{i\in\mathcal{I}}$ consists of the hyperplane $\{x \in \mathbb{R}^n : \langle x, w\rangle = 1\}$ with label $y = 1$ and the hyperplane $\{x \in \mathbb{R}^n : \langle x, w\rangle = 2\}$ with label $y = 2$. The solution to projection factor recovery in this case is $\Pi = ww^\top$. To see what NuMax delivers, observe that $\mathcal{S}_1(\mathcal{D})$ consists of all members of the unit sphere in $\mathbb{R}^n$ that are not orthogonal to w, and so the constraint $v^\top M v \geq 1-\delta$ for $v \in \mathcal{S}_1(\mathcal{D})$ is equivalent to $M \succeq (1-\delta)I$. It follows that $M = (1-\delta)I$ is the unique solution to NuMax($\mathcal{D},\delta$). On the other hand, SqueezeFit determines Π: For $\Delta \leq 1$, the unique solution to sqz($\mathcal{D},\Delta$) is $\Delta^2\Pi$ as a consequence of Theorem 11, Lemma 10(i), and Theorem 3, while sqz($\mathcal{D},\Delta$) is infeasible for $\Delta > 1$.

1.2 Notation

Given an objective function $f\colon \mathbb{R}^n \to \mathbb{R}$ and feasibility region $S \subseteq \mathbb{R}^n$, consider the program

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad x \in S \tag{P}$$

Then we denote

$$\operatorname{val} P := \inf_{x\in S} f(x), \qquad \arg P := \{x \in S : x = \operatorname{val} P\}.$$

If instead P were a maximization program, then we take $\operatorname{val} P := \sup_{x\in S} f(x)$. Given $A \in \mathbb{R}^{m\times n}$, we write $\operatorname{im}(A) := \{Ax : x \in \mathbb{R}^n\}$ to denote the image of A as a linear transformation. By default, all vectors are to be interpreted as column vectors. Throughout, 1 denotes an all-ones (column) vector, the dimensionality of which will be clear from context. For example, if $x \in \mathbb{R}^n$, then $1^\top x$ denotes the sum of the entries of x. Given a set T and $k \in \mathbb{N}$, we denote $\binom{T}{k} := \{S \subseteq T : |S| = k\}$. Given a function $f(k, n)$, we write $f(k, n) = O_k(g(n))$ if for every k, there exists a constant $C = C(k)$ such that $f(k, n) \leq C \cdot g(n)$ for every n. We also write $f(k, n) = o_{n\to\infty}(1)$ if for every k, it holds that $\lim_{n\to\infty} f(k, n) = 0$.

1.3 Outline

In the next section, we report conditions under which SqueezeFit successfully performs projection factor recovery. We prove these results in Section 3, where (as prerequisites) we also study fundamental geometric features of SqueezeFit, and then use these features to analyze strong duality. SqueezeFit also performs well in practice, which we illustrate in Section 4 with an assortment of numerical experiments. We conclude in Section 5 with a discussion of various open questions and opportunities for future work.

2 Projection factor recovery with SqueezeFit

Throughout, $\mathcal{D} = \{(x_i, y_i)\}_{i\in\mathcal{I}}$ denotes a sequence of points in $\mathbb{R}^d \times [k]$ with a possibly uncountable index set $\mathcal{I}$. We observe that SqueezeFit frequently succeeds in projection factor recovery when the

T-component $\{(\Pi x_i, y_i)\}_{i\in\mathcal{I}}$ of the data is "well behaved" and the $T^\perp$-component $\{(I-\Pi)x_i\}_{i\in\mathcal{I}}$ of the data is "independent" of the T-component. See Figure 1 for an illustrative example; in this case, we defined $\mathcal{D}$ by first selecting points from concentric circles in the xy-plane (the T-component) and then adding Gaussian noise in the z-direction (the $T^\perp$-component). In this section, we present two general instances of this phenomenon. The first instance enjoys a short proof:

Theorem 3. *Given $\mathcal{D}_0 = \{(x_i, y_i)\}_{i\in\mathcal{I}}$, select any nonempty set $\mathcal{S} \subseteq (\operatorname{span}\{x_i\}_{i\in\mathcal{I}})^\perp$ and define $\mathcal{D} = \{(x_i + s, y_i)\}_{i\in\mathcal{I}, s\in\mathcal{S}}$. Then*

(i) $\arg\operatorname{sqz}(\mathcal{D}, \Delta) = \arg\operatorname{sqz}(\mathcal{D}_0, \Delta)$*, and*

(ii) every $M \in \arg\operatorname{sqz}(\mathcal{D}_0, \Delta)$ *satisfies* $\operatorname{im}(M) \subseteq \operatorname{span}\{x_i\}_{i\in\mathcal{I}}$.

In particular, if the T-component $\mathcal{D}_0$ is "well behaved" (in the sense that every $M \in \arg\operatorname{sqz}(\mathcal{D}_0, \Delta)$ satisfies $\operatorname{im}(M) = \operatorname{span}\{x_i\}_{i\in\mathcal{I}}$), then SqueezeFit succeeds in projection factor recovery from $\mathcal{D}$ (just find any $M \in \arg\operatorname{sqz}(\mathcal{D}, \Delta)$ and take $\Pi = M(M^\top M)^{-1}M^\top$).

Proof of Theorem 3. We prove (ii) first. Let Π denote orthogonal projection onto $\operatorname{span}\{x_i\}_{i\in\mathcal{I}}$. Then for every M that is feasible in $\operatorname{sqz}(\mathcal{D}_0, \Delta)$, $\Pi M\Pi$ is also feasible with $\operatorname{tr}(\Pi M\Pi) = \operatorname{tr}(M\Pi) \leq \operatorname{tr} M$. The last inequality follows from the von Neumann trace inequality, where equality occurs only if $\operatorname{im}(M) \subseteq \operatorname{im}(\Pi)$. As such, $M \in \arg\operatorname{sqz}(\mathcal{D}_0, \Delta)$ only if $M = \Pi M\Pi$. For (i), $\mathcal{Z}(\mathcal{D}_0) \subseteq \mathcal{Z}(\mathcal{D})$, and so $\operatorname{sqz}(\mathcal{D}_0, \Delta)$ is a relaxation of $\operatorname{sqz}(\mathcal{D}, \Delta)$. Since every $\Pi M\Pi \in \arg\operatorname{sqz}(\mathcal{D}_0, \Delta)$ is trivially feasible in $\operatorname{sqz}(\mathcal{D}, \Delta)$, we then have $\arg\operatorname{sqz}(\mathcal{D}_0, \Delta) \subseteq \arg\operatorname{sqz}(\mathcal{D}, \Delta) \subseteq \arg\operatorname{sqz}(\mathcal{D}_0, \Delta)$. □

The above guarantee uses a weak notion of "well behaved" for the T-component of $\mathcal{D}$, but a strong notion of "independent" for the $T^\perp$-component. In what follows, we strengthen "well behaved" to mean Δ-*fixed* (defined below), and weaken "independent" so that the $T^\perp$-component isn't identical (but follows the same Gaussian distribution) as you vary the T-component.

Definition 4.

(i) The **contact vectors** of $\mathcal{D}$ are the shortest vectors in $\mathcal{Z}(\mathcal{D})$, when they exist.

(ii) We say $\mathcal{D} = \{(x_i, y_i)\}_{i\in\mathcal{I}}$ is **Δ-fixed** if there exists $M \in \arg\operatorname{sqz}(\mathcal{D}, \Delta)$ such that $M^{1/2}x_i = x_i$ for every $i \in \mathcal{I}$.

Intuitively, the contact vectors dictate which directions of the dataset are able to be squeezed. If the contact vectors are longer than Δ, then we can squeeze the data in all directions. If the contact vectors have length Δ, then we can only squeeze in the orthogonal complement of the contact vectors. If the contact vectors have length Δ and they span the data, then the data cannot be squeezed any further, and so the data is Δ-fixed. If the contact vectors have length smaller than Δ, then the SqueezeFit program is infeasible. We provide a rigorous treatment of these claims in Subsection 3.1. As an example of a Δ-fixed dataset, consider the far-right panel of Figure 1. Here, the classes form two concentric circles in the xy-plane; if we let Δ denote the distance between these two circles (i.e., the difference between their radii), then this dataset is Δ-fixed.

Definition 5. Given $\mathcal{D}_0 = \{(x_i, y_i)\}_{i\in[a]}$ in $\mathbb{R}^d \times [k]$, let Π denote orthogonal projection onto the r-dimensional $\operatorname{span}\{x_i\}_{i\in[a]}$ in $\mathbb{R}^d$. Select any $\sigma > 0$.

(i) For each $i \in [a]$, draw $\{g_{it}\}_{t\in[b]}$ independently from $\mathcal{N}(0, \sigma^2(I-\Pi))$, and consider the perturbed dataset $\mathcal{D} := \{(x_i + g_{it}, y_i)\}_{i\in[a], t\in[b]}$. We say $\mathcal{D}$ is drawn from the **projection factor model**, and we write $\mathcal{D} \sim \operatorname{PFM}(\mathcal{D}_0, \sigma^2, b)$.

(ii) Let $\mathscr{Z}_0$ denote the contact vectors of $\mathcal{D}_0$, and let λ denote the smallest non-zero eigenvalue of $\sum_{z\in\mathscr{Z}_0} zz^\top$. Then we write $\mathsf{SNR}(\mathcal{D}_0,\sigma^2):=\lambda/(2r\sigma^2)$.

We note that the data in the far-left panel of Figure 1 was drawn from $\mathrm{PFM}(\mathcal{D}_0,\sigma^2,1)$, where $\mathcal{D}_0$ corresponds to data points sampled along concentric circles in the xy-plane. In order to appreciate the above definition of SNR, first note that the contact vectors of $\mathcal{D}_0$ contain whatever "signal" SqueezeFit uses to find Π. In the idealized setting where the contact vectors consist of an orthogonal basis B for $T=\operatorname{span}\{x_i\}_{i\in[a]}$ together with its negation $-B$, then $\lambda/2$ equals the squared length Δ^2 of each contact vector. Since this energy is spread over r dimensions, we can say that the amount of signal per dimension is $\lambda/(2r)$. Intuitively, if the contact vectors "barely" span (meaning λ is small), then the signal is weaker, whereas additional contact vectors provide stronger signal. Our notion of signal-to-noise ratio SNR compares the amount of signal per dimension of T to the amount of noise per dimension of $T^\perp$.

Our main result requires a technical lemma, which in turn requires a definition: Given a closed convex cone $\mathcal{C}\subseteq\mathbb{R}^n$, the **statistical dimension** of $\mathcal{C}$ is given by

$$\delta(\mathcal{C}):=\mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,I)}\Big(\sup_{x\in\mathcal{C}\cap\mathbb{B}^n}\langle x,g\rangle\Big)^2,$$

where $\mathbb{B}^n$ denotes the unit Euclidean ball in $\mathbb{R}^n$. The notion of statistical dimension was introduced in [1] to characterize phase transitions in compressed sensing and elsewhere.

Lemma 6. *There exist universal constants $c_1,c_2,c_3>0$ for which the closed convex cone*

$$\mathcal{C}_n\;:=\;\Big\{\;v\in\mathbb{R}^n\;:\;v\geq 0,\;\;\max v\leq\frac{c_1\sqrt{\log n}}{n}\cdot 1^\top v\;\Big\}$$

has statistical dimension $\delta(\mathcal{C}_n)\in[c_2n,\tfrac{1}{2}n]$ for every $n\geq c_3$.

Theorem 7 (main result)**.** *Let $c_1,c_2,c_3>0$ be the constants in Lemma 6. Suppose $\mathcal{D}_0$ is Δ-fixed, and draw $\mathcal{D}\sim\mathrm{PFM}(\mathcal{D}_0,\sigma^2,b)$. Then $\arg\operatorname{sqz}(\mathcal{D},\Delta)=\{\Pi\}$ with probability at least $1-6|\mathscr{Z}_0|e^{-c_2b/48}$ provided*

$$b\geq\max\Big\{\frac{2}{c_2}(d-r),c_3\Big\},\qquad\mathsf{SNR}(\mathcal{D}_0,\sigma^2)\geq 3c_1\Big(4+\frac{c_2}{2}\Big)\sqrt{\log b}.$$

Our proof of Theorem 7 appears in Subsection 3.3 and leverages the following:

(i) $\arg\operatorname{sqz}(\mathcal{D}_0,\Delta)=\{\Pi\}$ because $\mathcal{D}_0$ is Δ-fixed,

(ii) the optimality of Π in $\arg\operatorname{sqz}(\mathcal{D}_0,\Delta)$ enjoys an easy-to-construct dual certificate, and

(iii) the dual certificate for $\mathcal{D}$ is a predictable perturbation of the dual certificate for $\mathcal{D}_0$.

We establish (i) in Subsection 3.1, (ii) in Subsection 3.2, and (iii) in Subsection 3.3. Interestingly, the SNR threshold in Theorem 7 is tight up to logarithmic factors:

Theorem 8. *Fix $k=2$ and $\Delta=1$. For every $\epsilon\in(0,2)$, there exists $d_0\in\mathbb{N}$, along with a sequence $\{(\mathcal{D}_0)_d\}_{d\geq d_0}$ of Δ-fixed datasets in $\mathbb{R}^d\times[k]$ and a sequence $\{\sigma_d\}_{d\geq d_0}$ in $\mathbb{R}_{\geq 0}$ such that*

(i) $\mathsf{SNR}(\mathcal{D}_0,\sigma_d^2)=\epsilon$ for every $d\geq d_0$, and

(ii) $\mathcal{D}_d\sim\mathrm{PFM}((\mathcal{D}_0)_d,\sigma_d^2,b_d)$ satisfies $\arg\operatorname{sqz}(\mathcal{D}_d,\Delta)=\{\Pi_d\}$ with probability at least $1/2$ for each $d\geq d_0$ only if b_d grows superpolynomially with d.

Going the other direction, as a consequence of Theorem 7, it holds that for every $\sigma > 0$, there exists a sufficiently large b such that $\arg\operatorname{sqz}(\mathcal{D}, \Delta) = \{\Pi\}$ with high probability:

Corollary 9. *Fix any Δ-fixed $\mathcal{D}_0$ and $\sigma > 0$. For each $b \in \mathbb{N}$, draw $\mathcal{D}_b \sim \mathrm{PFM}(\mathcal{D}_0, \sigma^2, b)$. Then*

$$\lim_{b\to\infty} \mathbb{P}\Big\{ \arg\operatorname{sqz}(\mathcal{D}_b, \Delta) \neq \{\Pi\} \Big\} = 0.$$

We note that for large σ, the rate of convergence in Corollary 9 is very slow, as we take b to be exponentially large in d in order to obtain $\arg\operatorname{sqz}(\mathcal{D}_b, \Delta) = \{\Pi\}$ with high probability. By contrast, such a large choice of b is unnecessary for projection factor recovery to be information theoretically possible. For example, it suffices to have $b > d - r$, even when σ is arbitrarily large. Indeed, for such b, it is straightforward to show that with probability 1, every size-b subcollection of $\{x_i + g_{it}\}_{i\in[a],t\in[b]}$ has affine rank $\geq d-r$, and the subcollections of affine rank $d-r$ are precisely those of the form $\{x_i + g_{it} : t \in [b]\}$. As such, for projection factor recovery, it suffices to first find the unique balanced partition of the data that minimizes maximum affine rank, then apply principal component analysis to one of the resulting size-b subcollections to recover $(\operatorname{span}\{x_i\}_{i\in[a]})^\perp$, and finally take Π to be orthogonal projection onto the orthogonal complement of this subspace. Interestingly, this method does not use the labels $\{y_i\}_{i\in[a]}$ to recover the projection. Of course, this procedure is not computationally tractable, and it heavily exploits the model of the data.

3 A theoretical treatment of SqueezeFit

3.1 The geometry of SqueezeFit

In this subsection, we rigorously treat the geometric intuition captured in the paragraph following Definition 5. First, we prove that Δ-fixed data enjoys a unique SqueezeFit optimizer, namely, the orthogonal projection onto the span of the data:

Lemma 10. *Pick any $\mathcal{D} = \{(x_i, y_i)\}_{i\in\mathcal{I}}$.*

(i) $\mathcal{D}$ is Δ-fixed if and only if the orthogonal projection onto $\operatorname{span}\{x_i\}_{i\in\mathcal{I}}$ is the unique member of $\arg\operatorname{sqz}(\mathcal{D}, \Delta)$.

(ii) If $\mathcal{D}$ is Δ-fixed, then $\operatorname{span}\mathcal{Z}(\mathcal{D}) = \operatorname{span}\{x_i\}_{i\in\mathcal{I}}$.

Proof. (i) First, $(\Leftarrow)$ is immediate. For $(\Rightarrow)$, we have by assumption that there exists $M \in \arg\operatorname{sqz}(\mathcal{D}, \Delta)$ with a leading eigenvalue of 1 whose eigenspace contains every x_i. Let Π denote orthogonal projection onto $\operatorname{span}\{x_i\}_{i\in\mathcal{I}}$. Then we may write $M = \Pi + \Gamma$ for some Γ satisfying $0 \preceq \Gamma \preceq I$ and $\Gamma\Pi = \Pi\Gamma = 0$. Note that Π is feasible in $\operatorname{sqz}(\mathcal{D}, \Delta)$ since $0 \preceq \Pi \preceq I$ and

$$z^\top \Pi z = (\Pi z)^\top (\Pi z) = z^\top z \geq z^\top M z \geq \Delta^2$$

for every $z \in \mathcal{Z}(\mathcal{D})$. Finally, we must have $M = \Pi$, i.e., $\Gamma = 0$, since otherwise $\operatorname{tr}\Pi < \operatorname{tr}\Pi + \operatorname{tr}\Gamma = \operatorname{tr} M$, thereby violating the assumption that $M \in \arg\operatorname{sqz}(\mathcal{D}, \Delta)$.

(ii) By (i), Π is feasible in $\operatorname{sqz}(\mathcal{D}, \Delta)$. Let $\Pi_\mathcal{Z}$ denote orthogonal projection onto $\operatorname{span}\mathcal{Z}(\mathcal{D})$. Then $0 \preceq \Pi_\mathcal{Z} \preceq I$ and

$$z^\top \Pi_\mathcal{Z} z = (\Pi_\mathcal{Z} z)^\top (\Pi_\mathcal{Z} z) = z^\top z \geq z^\top \Pi z \geq \Delta^2$$

for every $z \in \mathcal{Z}(\mathcal{D})$, and so $\Pi_\mathcal{Z}$ is also feasible in $\operatorname{sqz}(\mathcal{D}, \Delta)$. Since $\Pi \in \arg\operatorname{sqz}(\mathcal{D}, \Delta)$ by (i), we then have

$$\dim\operatorname{span}\{x_i\}_{i\in\mathcal{I}} = \operatorname{tr}\Pi \leq \operatorname{tr}\Pi_\mathcal{Z} = \dim\operatorname{span}\mathcal{Z}(\mathcal{D}).$$

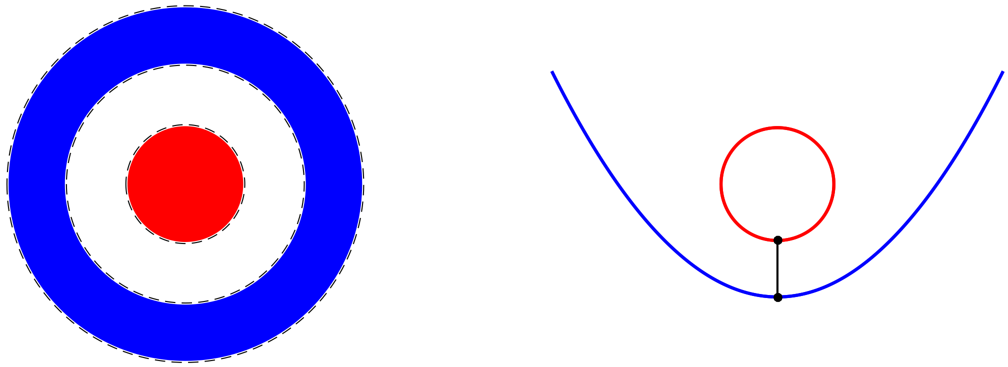

Figure 2: Examples of Δ-fixed data that fail to admit a spanning set of contact vectors. Both examples exhibit $k = 2$ classes in $\mathbb{R}^2$. On the left, the classes are open sets, and they admit no contact vectors. On the right, the classes are compact sets, and the only contact vectors are $(0, \pm 1)$. These examples illustrate why Theorem 11 requires $|\mathcal{Z}(\mathcal{D})| < \infty$ for the converse to hold. By Theorem 15, SqueezeFit fails to admit a Haar dual certificate in these examples.

The definition of $\mathcal{Z}(\mathcal{D})$ implies that $\operatorname{span} \mathcal{Z}(\mathcal{D}) \subseteq \operatorname{span}\{x_i\}_{i\in\mathcal{I}}$, and so the above dimension count gives the desired equality. $\square$

While Lemma 10(i) is a critical to our the proof of Theorem 7, Lemma 10(ii) will allow us to prove another portion of our geometric intuition: If the contact vectors have length Δ and they span the data, then the data cannot be squeezed any further, and so the data is Δ-fixed. We make this rigorous in the following:

Theorem 11. *If $\mathcal{D} = \{(x_i, y_i)\}_{i\in\mathcal{I}}$ has contact vectors of length Δ that span $\operatorname{span}\{x_i\}_{i\in\mathcal{I}}$, then $\mathcal{D}$ is Δ-fixed. Furthermore, the converse holds when $|\mathcal{Z}(\mathcal{D})| < \infty$.*

See Figure 2 for (necessarily infinite) examples in which the converse fails to hold. The proof of Theorem 11 requires the following lemma:

Lemma 12. *If $\mathcal{D}$ is Δ-fixed, then its contact vectors have length Δ, provided they exist.*

Proof. Suppose $\mathcal{D}$ has a contact vector. Since $\mathcal{D}$ is Δ-fixed, $\operatorname{sqz}(\mathcal{D}, \Delta)$ is feasible, and so this contact vector has length $\geq \Delta$. Suppose the contact vector has length $\Delta_0 > \Delta$, put $\alpha = (\Delta/\Delta_0)^2 < 1$, and select $\Pi \in \arg\operatorname{sqz}(\mathcal{D}, \Delta)$ such that $\Pi^{1/2}x_i = x_i$ for every $i \in \mathcal{I}$. By Lemma 10(i), Π is orthogonal projection onto $\operatorname{span}\{x_i\}_{i\in\mathcal{I}}$. We will show that $M := \alpha\Pi$ is feasible in $\operatorname{sqz}(\mathcal{D}, \Delta)$ with smaller trace than Π, contradicting the fact that Π is optimal in $\operatorname{sqz}(\mathcal{D}, \Delta)$. Since $\alpha \in (0, 1)$, we have $0 \preceq M \preceq I$. Next, every $z \in \mathcal{Z}(\mathcal{D})$ satisfies

$$z^\top M z = \alpha z^\top \Pi z = \alpha z^\top z \geq \Delta^2,$$

where the last equality applies Lemma 10(i) and the inequality uses the definition of α. Finally, $\operatorname{tr} M = \alpha \operatorname{tr} \Pi < \operatorname{tr} \Pi$, producing the desired contradiction. $\square$

Proof of Theorem 11. Pick any $M \in \arg\operatorname{sqz}(\mathcal{D}, \Delta)$. Then for every contact vector z of $\mathcal{D}$, we have

$$\Delta^2 \leq z^\top M z \leq z^\top z = \Delta^2.$$

Considering the far left- and right-hand sides, all inequalities are necessarily equalities. In particular, equality in the second inequality combined with $M \preceq I$ implies that M has a leading eigenvalue of 1 whose eigenspace contains every contact vector, and therefore every x_i (by assumption). This implies that $\mathcal{D}$ satisfies the definition of Δ-fixed.

For the converse, suppose $|\mathcal{Z}(\mathcal{D})| < \infty$ and $\mathcal{D}$ is Δ-fixed. Since $|\mathcal{Z}(\mathcal{D})| < \infty$, $\mathcal{D}$ necessarily has a contact vector. By Lemma 12, the contact vectors of $\mathcal{D}$ necessarily have length Δ. Let S denote the span of these contact vectors. We seek to prove $S = \operatorname{span}\{x_i\}_{i \in \mathcal{I}}$, and by Lemma 10(ii), it is equivalent to show $S = \operatorname{span} \mathcal{Z}(\mathcal{D})$. Since $S \subseteq \operatorname{span} \mathcal{Z}(\mathcal{D})$, it suffices to show $\operatorname{span} \mathcal{Z}(\mathcal{D}) \subseteq S$. To this end, consider the set $A := \{\|z\| : z \in \mathcal{Z}(\mathcal{D}) \setminus S\} \subseteq (\Delta, \infty)$. If A is empty, then we are done, since this implies $\mathcal{Z}(\mathcal{D}) \subseteq S$. Otherwise, since $|\mathcal{Z}(\mathcal{D})| < \infty$, we have $\alpha := (\Delta/(\min A))^2 < 1$. Let Π_S and Π_T denote orthogonal projection onto S and $T := S^\perp \cap \operatorname{span}\{x_i\}_{i \in \mathcal{I}}$, respectively, and select $\Pi \in \arg \operatorname{sqz}(\mathcal{D}, \Delta)$. We will show that $M := \Pi_S + \alpha \Pi_T$ is feasible in $\operatorname{sqz}(\mathcal{D}, \Delta)$ with smaller trace than Π, contradicting the fact that Π is optimal in $\operatorname{sqz}(\mathcal{D}, \Delta)$.

Since $\alpha \in (0, 1)$, we have $0 \preceq M \preceq I$. Next, we will verify $z^\top M z \geq \Delta^2$ for every $z \in \mathcal{Z}(\mathcal{D})$. For $z \in \mathcal{Z}(\mathcal{D}) \cap S$, we have

$$z^\top M z = z^\top (\Pi_S + \alpha \Pi_T) z = z^\top \Pi_S z = z^\top z \geq \Delta^2.$$

Meanwhile, for $z \in \mathcal{Z}(\mathcal{D}) \setminus S$, we have

$$z^\top M z = z^\top (\Pi_S + \alpha \Pi_T) z \geq \alpha z^\top (\Pi_S + \Pi_T) z \overset{\text{(a)}}{=} \alpha z^\top \Pi z = \alpha \|\Pi z\|^2 \overset{\text{(b)}}{=} \alpha \|z\|^2 \geq \Delta^2,$$

where (a) and (b) follow from Lemma 10(i), and the final inequality follows from the definition of α. Overall, M is feasible in $\operatorname{sqz}(\mathcal{D}, \Delta)$, and yet

$$\operatorname{tr} M = \operatorname{tr} \Pi_S + \alpha \operatorname{tr} \Pi_T < \operatorname{tr} \Pi_S + \operatorname{tr} \Pi_T = \operatorname{tr}(\Pi_S + \Pi_T) = \operatorname{tr} \Pi,$$

where the last equality again follows from Lemma 10(i). This is the desired contradiction. $\square$

As an aside, we establish that you need not squeeze data more than once:

Theorem 13. *Given $\mathcal{D} = \{(x_i, y_i)\}_{i \in \mathcal{I}}$, then $\{(M^{1/2} x_i, y_i)\}_{i \in \mathcal{I}}$ is Δ-fixed for every $M \in \arg \operatorname{sqz}(\mathcal{D}, \Delta)$.*

Proof. Pick $M \in \arg \operatorname{sqz}(\mathcal{D}, \Delta)$, put $\mathcal{D}' = \{(M^{1/2} x_i, y_i)\}_{i \in \mathcal{I}}$, and pick $N \in \arg \operatorname{sqz}(\mathcal{D}', \Delta)$. We claim that $M_0 := (M^{1/2})^\top N M^{1/2}$ is feasible in $\operatorname{sqz}(\mathcal{D}, \Delta)$. First, we have $\mathcal{Z}(\mathcal{D}') = M^{1/2} \mathcal{Z}(\mathcal{D})$, and so the feasibility of N in $\operatorname{sqz}(\mathcal{D}', \Delta)$ implies

$$z^\top M_0 z = (M^{1/2} z)^\top N (M^{1/2} z) \geq \Delta^2$$

for every $z \in \mathcal{Z}(\mathcal{D})$. Similarly, $0 \preceq M_0 \preceq I$ follows from the facts that $0 \preceq N \preceq I$ and $M \preceq I$:

$$x^\top M_0 x = (M^{1/2} x)^\top N (M^{1/2} x) \geq 0,$$
$$x^\top M_0 x = (M^{1/2} x)^\top N (M^{1/2} x) \leq (M^{1/2} x)^\top (M^{1/2} x) = x^\top M x \leq x^\top x$$

for every $x \in \mathbb{R}^d$. Overall, we indeed have that M_0 is feasible in $\operatorname{sqz}(\mathcal{D}, \Delta)$.

Next, let $\alpha_1 \geq \cdots \geq \alpha_d$ and $\beta_1 \geq \cdots \geq \beta_d$ denote the eigenvalues of M and N, respectively. Then the von Neumann trace inequality gives

$$\sum_{j=1}^d \alpha_j = \operatorname{tr} M \leq \operatorname{tr} M_0 = \operatorname{tr}(MN) \leq \sum_{j=1}^d \alpha_j \beta_j \leq \sum_{j=1}^d \alpha_j,$$

where the last inequality uses the facts that $\alpha_j \geq 0$ and $\beta_j \leq 1$ for every j. Considering the far left- and right-hand sides, all inequalities are necessarily equalities. Equality in the von Neumann trace inequality implies that M and N are simultaneously unitarily diagonalizable, while equality in the last inequality implies that $\beta_j = 1$ whenever $\alpha_j \neq 0$. As such, $N^{1/2}$ fixes the column space of $M^{1/2}$, meaning $\mathcal{D}' = \{(M^{1/2} x_i, y_i)\}_{i \in \mathcal{I}}$ satisfies the definition of Δ-fixed, as desired. $\square$

3.2 Conditions for strong duality

Our proof of Theorem 7 leverages the construction of a dual certificate, motivating our study of the dual program. We follow [16] to find the Haar dual program of $\mathrm{sqz}(\mathcal{D}, \Delta)$ in the case where $\mathcal{Z}(\mathcal{D})$ is infinite:

$$\begin{aligned} \sup \quad & \Delta^2 \sum_{z \in \mathcal{Z}(\mathcal{D})} \gamma(z) - \operatorname{tr} Y \\ \text{subject to} \quad & \sum_{z \in \mathcal{Z}(\mathcal{D})} \gamma(z) z z^\top - Y \preceq I, \quad Y \succeq 0, \quad \gamma \geq 0, \quad |\operatorname{supp}(\gamma)| < \infty \end{aligned} \qquad (\mathrm{dual}(\mathcal{D}, \Delta))$$

Here, the decision variables are $\gamma \colon \mathcal{Z}(\mathcal{D}) \to \mathbb{R}$ and $Y \in \mathbb{R}^{d \times d}$. The above program reduces to the dual semidefinite program when $\mathcal{Z}(\mathcal{D})$ is finite. In either case, weak duality gives that $\operatorname{val}\mathrm{sqz}(\mathcal{D}, \Delta) \geq \operatorname{val}\mathrm{dual}(\mathcal{D}, \Delta)$ when $\mathrm{sqz}(\mathcal{D}, \Delta)$ is feasible. We are interested in when $\mathrm{sqz}(\mathcal{D}, \Delta)$ admits a **Haar dual certificate**, that is, a maximizer of $\mathrm{dual}(\mathcal{D}, \Delta)$. Indeed, a Haar dual certificate certifies the optimality of a given optimal point in $\mathrm{sqz}(\mathcal{D}, \Delta)$ by witnessing equality in weak duality. In the finite case, we can prove the existence of a Haar dual certificate by manipulating $\mathrm{sqz}(\mathcal{D}, \Delta)$ and applying Slater's condition:

Theorem 14. *Suppose $|\mathcal{Z}(\mathcal{D})| < \infty$ and $\mathrm{sqz}(\mathcal{D}, \Delta)$ is feasible. Then $\mathrm{sqz}(\mathcal{D}, \Delta)$ admits a Haar dual certificate.*

Proof. Put $\mathcal{Z}_1 = \{z \in \mathcal{Z}(\mathcal{D}) : \|z\| = \Delta\}$, $\mathcal{Z}_2 = \mathcal{Z}(\mathcal{D}) \setminus \mathcal{Z}_1$, and $T = \operatorname{span}\{z\}_{z \in \mathcal{Z}_1}$, and consider the related semidefinite program:

$$\begin{aligned} \min \quad & \operatorname{tr}(\Pi_{T^\perp} X \Pi_{T^\perp}) \\ \text{subject to} \quad & z^\top \Pi_{T^\perp} X \Pi_{T^\perp} z \geq \Delta^2 - \|\Pi_T z\|^2 \;\; \forall z \in \mathcal{Z}_2, \quad \Pi_T X \Pi_T = 0, \quad 0 \preceq X \preceq I \end{aligned} \qquad (2)$$

Importantly, every feasible point X of (2) can be transformed into a feasible point $M = \Pi_{T^\perp} X \Pi_{T^\perp} + \Pi_T$ of $\mathrm{sqz}(\mathcal{D}, \Delta)$. Indeed, if $z \in \mathcal{Z}_2$, then

$$z^\top (\Pi_{T^\perp} X \Pi_{T^\perp} + \Pi_T) z = z^\top \Pi_{T^\perp} X \Pi_{T^\perp} z + \|\Pi_T z\|^2 \geq \Delta^2,$$

while if $z \in \mathcal{Z}_1$, then $z \in T$, and so $z^\top (\Pi_{T^\perp} X \Pi_{T^\perp} + \Pi_T) z = \|z\|^2 = \Delta^2$. Furthermore, $0 \preceq \Pi_{T^\perp} X \Pi_{T^\perp} + \Pi_T \preceq \Pi_{T^\perp} + \Pi_T = I$.

Next, we demonstrate that the related program (2) satisfies strong duality by Slater's theorem. To see this, define $\alpha = \frac{1}{2} + \frac{1}{2\Delta^2} \min\{\|z\|^2 : z \in \mathcal{Z}_2\}$. Then $1 < \alpha < \|z\|^2/\Delta^2$ for every $z \in \mathcal{Z}_2$. Consider $X_0 = \frac{1}{\alpha} \Pi_{T^\perp}$. Then for every $z \in \mathcal{Z}_2$, we have

$$z^\top \Pi_{T^\perp} X_0 \Pi_{T^\perp} z = \frac{1}{\alpha} \|\Pi_{T^\perp} z\|^2 = \frac{1}{\alpha} \|z\|^2 - \frac{1}{\alpha} \|\Pi_T z\|^2 > \Delta^2 - \|\Pi_T z\|^2.$$

Furthermore, X_0 lies in the relative interior of $\{X : \Pi_T X \Pi_T = 0, 0 \preceq X \preceq I\}$. Overall, X_0 lies in the relative interior of the feasibility region of (2), and so Slater's theorem [6] implies that the value of (2) equals the value of its dual:

$$\begin{aligned} \max \quad & \sum_{z \in \mathcal{Z}_2} (\Delta^2 - \|\Pi_T z\|^2) \tilde{\gamma}(z) - \operatorname{tr} \tilde{Y} \\ \text{subject to} \quad & \sum_{z \in \mathcal{Z}_2} \tilde{\gamma}(z) (\Pi_{T^\perp} z)(\Pi_{T^\perp} z)^\top - \tilde{Y} \preceq \Pi_{T^\perp}, \quad \Pi_T \tilde{Y} \Pi_T = 0, \quad \tilde{Y} \succeq 0, \quad \tilde{\gamma} \geq 0 \end{aligned} \qquad (3)$$

Here, the decision variables are $\tilde{\gamma}\colon \mathcal{Z}_2 \to \mathbb{R}$ and $\tilde{Y} \in \mathbb{R}^{d\times d}$.

Next, we demonstrate how to transform every feasible point $(\tilde{\gamma}, \tilde{Y})$ of (3) into a feasible point (γ, Y) of $\operatorname{dual}(\mathcal{D},\Delta)$. To this end, define $Z = \sum_{z\in\mathcal{Z}_1} zz^\top$, and let λ denote the smallest nonzero eigenvalue of Z. Then we take

$$\gamma(z) := \left\{ \begin{array}{ll} 1/\lambda & \text{if } z \in \mathcal{Z}_1 \\ \tilde{\gamma}(z) & \text{if } z \in \mathcal{Z}_2 \end{array} \right\}, \qquad Y := \sum_{z\in\mathcal{Z}_2} \tilde{\gamma}(z)(\Pi_T z)(\Pi_T z)^\top + \frac{1}{\lambda} Z - \Pi_T + \tilde{Y}.$$

Then we immediately have $\gamma \geq 0$ and $Y \succeq 0$. Furthermore,

$$R := I - \sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z) zz^\top + Y = \Pi_{T^\perp} - \sum_{z\in\mathcal{Z}_2} \tilde{\gamma}(z)\Big(zz^\top - (\Pi_T z)(\Pi_T z)^\top\Big) + \tilde{Y}$$

satisfies $\Pi_T R \Pi_T = 0$ and $\Pi_{T^\perp} R \Pi_{T^\perp} \succeq 0$ by feasibility in (3), and so $R \succeq 0$, as desired.

By assumption, $\operatorname{sqz}(\mathcal{D},\Delta)$ is feasible, and so $\Pi_{T^\perp}$ is feasible in (2); also, $(\tilde{\gamma}, \tilde{Y}) = (0,0)$ is feasible in (3). By this feasibility and the extreme value theorem, we may take optimizers $X^\natural$ and $(\tilde{\gamma}^\natural, \tilde{Y}^\natural)$ of (2) and (3), respectively, and let M and (γ, Y) denote corresponding feasible points of $\operatorname{sqz}(\mathcal{D},\Delta)$ and $\operatorname{dual}(\mathcal{D},\Delta)$, respectively. Then the dual value of (γ, Y) equals the primal value of M:

$$\begin{aligned} \Delta^2 \sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z) - \operatorname{tr} Y &= \Delta^2 \sum_{z\in\mathcal{Z}_2} \tilde{\gamma}(z) + \frac{\Delta^2}{\lambda}|\mathcal{Z}_1| - \bigg(\sum_{z\in\mathcal{Z}_2} \tilde{\gamma}^\natural(z)\|\Pi_T z\|^2 + \frac{\Delta^2}{\lambda}|\mathcal{Z}_1| - \dim T + \operatorname{tr} \tilde{Y}^\natural \bigg) \\ &= \operatorname{val}(3) + \dim T = \operatorname{val}(2) + \dim T = \operatorname{tr}(\Pi_{T^\perp} X^\natural \Pi_{T^\perp}) + \operatorname{tr} \Pi_T = \operatorname{tr} M. \end{aligned}$$

Combining this with weak duality gives

$$\operatorname{val}\operatorname{dual}(\mathcal{D},\Delta) \geq \Delta^2 \sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z) - \operatorname{tr} Y = \operatorname{tr} M \geq \operatorname{val}\operatorname{sqz}(\mathcal{D},\Delta) \geq \operatorname{val}\operatorname{dual}(\mathcal{D},\Delta).$$

Considering the far left- and right-hand sides, we may conclude the desired equality. □

In the infinite case, strong duality is no longer guaranteed. Interestingly, we can characterize strong duality for Δ-fixed data:

Theorem 15. *Suppose $\mathcal{D} = \{(x_i, y_i)\}_{i\in\mathcal{I}}$ is Δ-fixed. Then $\operatorname{sqz}(\mathcal{D},\Delta)$ admits a Haar dual certificate if and only if the contact vectors of $\mathcal{D}$ span $\operatorname{span}\{x_i\}_{i\in\mathcal{I}}$.*

Proof. ($\Leftarrow$) Select any finite collection $\mathcal{Z}_0$ of contact vectors of $\mathcal{D}$ that span $\operatorname{span}\{x_i\}_{i\in\mathcal{I}}$, and put $X := \sum_{z\in\mathcal{Z}_0} zz^\top$. Let λ be the smallest non-zero eigenvalue of X, and define

$$\gamma(z) := \left\{ \begin{array}{ll} 1/\lambda & \text{if } z \in \mathcal{Z}_0 \\ 0 & \text{if } z \in \mathcal{Z}(\mathcal{D}) \setminus \mathcal{Z}_0 \end{array} \right\}, \qquad Y := \frac{1}{\lambda} X - \Pi,$$

where Π denotes orthogonal projection onto $\operatorname{span}\{x_i\}_{i\in\mathcal{I}}$. It is straightforward to check that (γ, Y) is feasible in $\operatorname{dual}(\mathcal{D},\Delta)$ with objective value $\operatorname{tr}\Pi$. By Lemma 10(i) and weak duality, (γ, Y) is therefore a Haar dual certificate.

($\Rightarrow$) Let Π_γ denote orthogonal projection onto the column space of $\sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z) zz^\top$. We first claim that if (γ, Y) is feasible in $\operatorname{dual}(\mathcal{D},\Delta)$, then so is $(\gamma, \Pi_\gamma Y \Pi_\gamma)$, and with monotonically larger objective value. Indeed, feasibility follows from

$$I - \sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z) zz^\top + \Pi_\gamma Y \Pi_\gamma = (I - \Pi_\gamma) + \Pi_\gamma\bigg(I - \sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z) zz^\top + Y\bigg)\Pi_\gamma \succeq 0,$$

and the objective value is monotonically larger since $\operatorname{tr}(\Pi_\gamma Y \Pi_\gamma) = \operatorname{tr}(Y\Pi_\gamma) \leq \operatorname{tr}(Y)$, where the last step follows from the von Neumann trace inequality. As such, the assumed Haar dual certificate (γ, Y) satisfies $\operatorname{im}(Y) \subseteq \operatorname{im}(\Pi_\gamma)$ without loss of generality.

Put $Q := \sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z)\frac{\Delta^2}{\|z\|^2}zz^\top - Y$. Then $\operatorname{tr} Q$ equals the dual value of (γ, Y). Let Π denote orthogonal projection onto $\operatorname{span}\{x_i\}_{i\in\mathcal{I}}$. Then $\operatorname{im}(Y) \subseteq \operatorname{im}(\Pi_\gamma) \subseteq \operatorname{im}(\Pi)$, and so we may strengthen an inequality that is implied by the dual feasibility of (γ, Y):

$$A := \Pi - \sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z)\Big(1 - \frac{\Delta^2}{\|z\|^2}\Big)zz^\top - Q = \Pi - \sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z)zz^\top + Y \succeq 0.$$

By Lemma 10(i), $\Pi \in \arg\operatorname{sqz}(\mathcal{D}, \Delta)$, and so $\operatorname{tr}\Pi = \operatorname{tr} Q$. As such,

$$0 \leq \operatorname{tr} A = -\sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z)(\|z\|^2 - \Delta^2) \leq 0.$$

Considering the far left- and right-hand sides, we infer two important conclusions:

(a) Since $A \succeq 0$ and $\operatorname{tr} A = 0$, we necessarily have $A = 0$.

(b) Since $\displaystyle\sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z)(\|z\|^2 - \Delta^2) = 0$, then $\gamma(z) > 0$ only if z is a contact vector of $\mathcal{D}$.

To be explicit, (b) applies Lemma 12. Let T denote the span of the contact vectors of $\mathcal{D}$. Rearranging $A = 0$ from (a) gives $\sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z)zz^\top = \Pi + Y$, and so (b) implies

$$\operatorname{span}\{x_i\}_{i\in\mathcal{I}} = \operatorname{im}(\Pi) \subseteq T \subseteq \operatorname{span}\{x_i\}_{i\in\mathcal{I}},$$

meaning $T = \operatorname{span}\{x_i\}_{i\in\mathcal{I}}$, as desired. □

When strong duality holds, one may seek a dual certificate of a given SqueezeFit optimizer. The following lemma facilitates this pursuit:

Lemma 16 (complementary slackness)**.** *Suppose* $\operatorname{sqz}(\mathcal{D}, \Delta)$ *admits a Haar dual certificate and select any* $M \in \arg\operatorname{sqz}(\mathcal{D}, \Delta)$*. Then* $\arg\operatorname{dual}(\mathcal{D}, \Delta)$ *is the set of points* (γ, Y) *that are feasible in* $\operatorname{dual}(\mathcal{D}, \Delta)$ *and further satisfy*

$$\operatorname{supp}(\gamma) \subseteq \Big\{z \in \mathcal{Z}(\mathcal{D}) : \|M^{1/2}z\| = \Delta\Big\}, \tag{4}$$

$$\operatorname{im}(Y) \subseteq \Big\{x : Mx = x\Big\}, \tag{5}$$

$$M = \sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z)(M^{1/2}z)(M^{1/2}z)^\top - Y. \tag{6}$$

Proof. Suppose (γ, Y) is feasible in $\operatorname{dual}(\mathcal{D}, \Delta)$. Then $I - \sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z)zz^\top + Y \succeq 0$. Multiplying by $M^{1/2}$ on both sides then gives

$$M - \sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z)M^{1/2}zz^\top M^{1/2} + M^{1/2}YM^{1/2} \succeq 0. \tag{7}$$

We take the trace and rearrange to get

$$\operatorname{tr} M \geq \sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z)z^\top Mz - \operatorname{tr}(MY) \geq \Delta^2 \sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z) - \operatorname{tr}(MY) \geq \Delta^2 \sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z) - \operatorname{tr} Y,$$

where the last step applied the von Neumann trace inequality. Since $\mathrm{sqz}(\mathcal{D},\Delta)$ admits a Haar dual certificate by assumption, $\arg\mathrm{dual}(\mathcal{D},\Delta)$ is the set of points (γ,Y) that are feasible in $\mathrm{dual}(\mathcal{D},\Delta)$ and further make all of the above inequalities achieve equality.

Equality in the second inequality is characterized by (4), while equality in the third inequality is characterized by (5). Equality in the first inequality occurs precisely when the positive-semidefinite matrix in (7) has trace zero, i.e., the matrix equals zero. As such,

$$M=\sum_{z\in\mathcal{Z}(\mathcal{D})}\gamma(z)M^{1/2}zz^\top M^{1/2}-M^{1/2}YM^{1/2}.$$

Furthermore, (5) implies $M^{1/2}YM^{1/2}=Y$, and so the above is equivalent to (6). □

In the finite case, strong duality is guaranteed by Theorem 14. Given $M\in\arg\mathrm{sqz}(\mathcal{D},\Delta)$, then Lemma 16 enables a quick procedure to find a dual certificate for M. Denote

$$\mathcal{Z}_0:=\Big\{z\in\mathcal{Z}(\mathcal{D}):\|M^{1/2}z\|=\Delta\Big\},\qquad E:=\Big\{x:Mx=x\Big\},$$

and consider the feasibility semidefinite program

$$\begin{aligned}\text{find}\quad&(\gamma,Y)\\ \text{subject to}\quad&\sum_{z\in\mathcal{Z}_0}\gamma(z)zz^\top-Y\preceq I,\quad \Pi_{E^\perp}Y\Pi_{E^\perp}=0,\\ &M=\sum_{z\in\mathcal{Z}_0}\gamma(z)(M^{1/2}z)(M^{1/2}z)^\top-Y,\quad Y\succeq 0,\quad \gamma\geq 0\end{aligned}\tag{8}$$

Importantly, solving (8) is much faster than solving $\mathrm{dual}(\mathcal{D},\Delta)$ since $|\mathcal{Z}_0|\ll|\mathcal{Z}(\mathcal{D})|$, and so (8) can be used to promote any heuristic SqueezeFit solver to a fast certifiably correct algorithm, much like [2, 20, 31]. Indeed, given a prospective solution M satisfying $0\preceq M\preceq I$, we may:

(i) Find the shortest vectors in $\{M^{1/2}z\}_{z\in\mathcal{Z}(\mathcal{D})}$ from $\{M^{1/2}x_i\}_{i\in\mathcal{I}}$. If these shortest vectors have length Δ, then this certifies that M is feasible in $\mathrm{sqz}(\mathcal{D},\Delta)$ and gives $\mathcal{Z}_0$ for the next step.

(ii) Solve the feasibility semidefinite program (8) to find a dual certificate (γ,Y).

By Lemma 16, the primal value of M equals the dual value of (γ,Y), and one may verify this *a posteriori*. Weak duality then implies $M\in\arg\mathrm{sqz}(\mathcal{D},\Delta)$.

One way to solve (i) is to partition $\mathcal{Z}(\mathcal{D})$ into subsets $\mathcal{Z}_{st}:=\{x_i-x_i:i,j\in\mathcal{I},y_i=s,y_j=t\}$ and then find the shortest vectors in $\{M^{1/2}z\}_{z\in\mathcal{Z}_{st}}$ from $\{M^{1/2}x_i\}_{i\in\mathcal{I}}$ for each $s,t\in[k]$ with $s\neq t$. This amounts to a fundamental problem in computational geometry: Given $A,B\in\mathbb{R}^d$, find the closest pairs $(x,y)\in A\times B$. Litiu and Kountanis [26] devised an $O_d(n\log^{d-1}n)$ divide-and-conquer algorithm that solves the problem for the taxicab metric in the special case where A and B are linearly separable. In practice, one might construct a k-d tree for A in $O_d(n\log n)$ time, and then use it to perform nearest neighbor search in $O_d(\log n)$ time on average [15] for each member of B. Next, (ii) is polynomial in d and $|\mathcal{Z}_0|$, and furthermore, Lemma 17 below gives that $|\mathcal{Z}_0|\leq d^4$ for generic data (we suspect this upper bound is loose). Overall, one may expect to accomplish (i) and (ii) in time that is roughly linear in n.

Lemma 17. *For every $d,k,n>0$ and every $\{y_i\}_{i\in[n]}\in[k]^n$, there exists a set $\mathcal{X}$ that is open and dense in $(\mathbb{R}^d)^n$ such that for every $\{x_i\}_{i\in[n]}\in\mathcal{X}$, $\Delta>0$, and every $M\in\arg\mathrm{sqz}(\{(x_i,y_i)\}_{i\in\mathcal{I}},\Delta)$,*

$$\Big|\Big\{(i,j):1\leq i<j\leq n,y_i\neq y_j,\|M^{1/2}(x_i-x_j)\|=\Delta\Big\}\Big|<\Big(\tbinom{d+1}{2}+1\Big)^2.$$

Proof. Fix $d, k, n > 0$ and $\{y_i\}_{i\in[n]} \in [k]^n$, and denote $\Omega := \{(i,j) : 1 \le i < j \le n, y_i \neq y_j\}$ and $D := \binom{d+1}{2}$. We may assume $|\Omega| \ge (D+1)^2$, since the result is otherwise immediate. In this case, we will find $\mathcal{X}$ for which something stronger than the desired conclusion holds: For every $\{x_i\}_{i\in[n]} \in \mathcal{X}$ and every $\mathcal{J} \in \binom{\Omega}{(D+1)^2}$, there is no nonzero $L \in \mathbb{R}^{d\times d}$ such that $\|L(x_i - x_j)\|^2$ is constant over $(i,j) \in \mathcal{J}$. In particular, for every such $\mathcal{J}$, we will find $\mathcal{K} \in \binom{\mathcal{J}}{D+1}$ for which there is no $L \in \mathbb{R}^{d\times d}$ such that $\|L(x_i - x_j)\|^2 = 1$ for every $(i,j) \in \mathcal{K}$.

We start by finding $\{z_l\}_{l\in[D+1]} \in (\mathbb{R}^d)^{D+1}$ for which there is no $L \in \mathbb{R}^{d\times d}$ such that $\|Lz_l\|^2 = 1$ for every $l \in [D+1]$. Let $\{A_i\}_{i\in[D]}$ be any basis of the D-dimensional vector space of symmetric matrices. By the spectral theorem, each A_i can be decomposed as a linear combination of rank-1 matrices $\{v_{ij}v_{ij}^\top\}_{j\in[d]}$ of unit trace, and so $\{v_{ij}v_{ij}^\top\}_{i\in[D],j\in[d]}$ spans the vector space. Select any basis from this spanning set, define the first D of the z_l's to be the corresponding v_{ij}'s, and take $z_{D+1} := 2z_1$. Since $\{z_l z_l^\top\}_{l\in[D]}$ is a basis of unit-trace matrices, we have that $1 = \|Lz_l\|^2 = \langle L^\top L, z_l z_l^\top\rangle$ for every $l \in [D]$ if and only if $L^\top L = I$. In this case, $\|Lz_{D+1}\|^2 = 4\|Lz_1\|^2 = 4$, and so there is no $L \in \mathbb{R}^{d\times d}$ such that $\|Lz_l\|^2 = 1$ for every $l \in [D+1]$.

Next, for each $\mathcal{K} \in \binom{\Omega}{D+1}$, we will construct a polynomial $p_\mathcal{K} \in \mathbb{R}[X_{ij} : i \in [d], j \in [n]]$ such that $p_\mathcal{K}(P)$ is nonzero only if there is no $L \in \mathbb{R}^{d\times d}$ such that $\|L(x_i - x_j)\|^2 = 1$ for every $(i,j) \in \mathcal{K}$. To this end, select any basis $\{A_l\}_{l\in[D]}$ of the D-dimensional vector space of symmetric matrices. For each $(i,j) \in \mathcal{K}$, consider the decomposition $(x_i - x_j)(x_i - x_j)^\top = \sum_{l\in[D]} c_{(i,j),l}A_l$, and let $F \in \mathbb{R}^{\mathcal{K}\times[D+1]}$ be defined by

$$F_{(i,j),l} = \begin{cases} c_{(i,j),l} & \text{if } l \in [D] \\ 1 & \text{if } l = D+1. \end{cases}$$

We will take $p_\mathcal{K}$ to be the polynomial that maps $\{x_i\}_{i\in[n]}$ to $\det(F)$. Indeed, if $\det(F) \neq 0$, then the all-ones vector does not lie in the span of the first D columns of F, i.e., there is no L such that

$$\|L(x_i - x_j)\|^2 = \langle L^\top L, (x_i - x_j)(x_i - x_j)^\top\rangle = \sum_{l\in[D]} c_{(i,j),l}\langle L^\top L, A_l\rangle = 1 \qquad \forall (i,j) \in \mathcal{K}.$$

We claim that $p_\mathcal{K}$ is a nonzero polynomial provided there exists $\{x_i\}_{i\in[n]} \in (\mathbb{R}^d)^n$ and a bijection $f : \mathcal{K} \to [D+1]$ such that $x_i - x_j = z_{f(i,j)}$, where $\{z_l\}_{l\in[D+1]}$ is the example constructed above. Indeed, since $\{z_l z_l^\top\}_{l\in[D]}$ is a basis, the corresponding $D \times D$ block of F has full rank, meaning the first D columns of F are linearly independent. By construction, these columns are also independent of the all-ones vector, and so $\det(F) \neq 0$. Finally, since there exists $\{x_i\}_{i\in[n]} \in (\mathbb{R}^d)^n$ such that $p_\mathcal{K}(\{x_i\}_{i\in[n]}) = \det(F) \neq 0$, we must have $p_\mathcal{K} \neq 0$.

We now use the polynomials $p_\mathcal{K}$ to construct a larger polynomial $p \in \mathbb{R}[X_{ij} : i \in [d], j \in [n]]$:

$$p(X) := \prod_{\mathcal{J}\in\binom{\Omega}{(D+1)^2}} \sum_{\mathcal{K}\in\binom{\mathcal{J}}{D+1}} p_\mathcal{K}(X)^2.$$

We claim that the result follows by taking $\mathcal{X} = \{\{x_i\}_{i\in[n]} : p(\{x_i\}_{i\in[n]}) \neq 0\}$. By construction, we have that every $\{x_i\}_{i\in[n]} \in \mathcal{X}$ satisfies $p(\{x_i\}_{i\in[n]}) \neq 0$, meaning that for every $\mathcal{J} \in \binom{\Omega}{(D+1)^2}$, there exists $\mathcal{K} \in \binom{\mathcal{J}}{D+1}$ such that $p_\mathcal{K}(\{x_i\}_{i\in[n]}) \neq 0$ (implying there is no $L \in \mathbb{R}^{d\times d}$ such that $\|L(x_i - x_j)\|^2 = 1$ for every $(i,j) \in \mathcal{K}$). It remains to establish that $\mathcal{X}$ is open and dense, i.e., that $p \neq 0$, or equivalently, that for every $\mathcal{J} \in \binom{\Omega}{(D+1)^2}$, there exists $\mathcal{K} \in \binom{\mathcal{J}}{D+1}$ such that $p_\mathcal{K} \neq 0$.

Pick $\mathcal{J} \in \binom{\Omega}{(D+1)^2}$ and consider the graph with vertex set $[n]$ and edge set $\mathcal{J}$. This graph has $(D+1)^2$ edges, and so the main result in [17] implies that either (i) there exists a vertex of degree at least $D+1$, or (ii) there exists a matching of size at least $D+1$. In the case of (i), let $\mathcal{K}$ be any $D+1$

of the edges incident to the vertex of maximum degree. Then $\mathcal{K}=\{(i,j):i\in\mathcal{K}'\}$ for some $j\in[n]$ and $\mathcal{K}'\in\binom{[n]\setminus\{j\}}{D+1}$. In this case, we can take any bijection $f'\colon\mathcal{K}'\to[D+1]$ and define $\{x_i\}_{i\in[n]}$ so that $x_i=z_{f'(i)}$ whenever $i\in\mathcal{K}'$ and otherwise $x_i=0$ (here, z_l is defined in the example above). Then $f\colon\mathcal{K}\to[D+1]$ defined by $f(i,j)=f'(i)$ is also a bijection and $x_i-x_j=x_i=z_{f'(i)}=z_{f(i,j)}$ for every $(i,j)\in\mathcal{K}$, implying $p_{\mathcal{K}}\neq 0$. In the case of (ii), let $\mathcal{K}$ be any $D+1$ of the edges in the maximum matching, and define $\{x_i\}_{i\in[n]}$ as follows: Select any bijection $f\colon\mathcal{K}\to[D+1]$, and for each $(i,j)\in\mathcal{K}$, define $x_i=z_{f(i,j)}$ and $x_j=0$. (For each vertex $i\in[n]$ that is not incident to an edge in $\mathcal{K}$, we may take $x_i=0$, say.) Then $x_i-x_j=x_i=z_{f(i,j)}$ for every $(i,j)\in\mathcal{K}$, and so $p_{\mathcal{K}}\neq 0$. In either case, we have that there exists $\mathcal{K}\in\binom{\mathcal{J}}{D+1}$ such that $p_{\mathcal{K}}\neq 0$, as desired. □

3.3 Proofs for Section 2

Having studied the geometric features and duality theory of SqueezeFit, we are now ready to prove the results from Section 2.

Proof of Theorem 8. Fix $\epsilon\in(0,2)$. We will show that for every $\alpha\in(0,(1+\frac{\epsilon}{2})^{-1})$, there exists $d_0=d_0(\alpha)$ such that for every $d\geq d_0$, there exists a requisite Δ-fixed $\mathcal{D}_0$ with $r:=\operatorname{rank}(\Pi)=\lceil\alpha d\rceil$. To see this, let $\{e_i\}_{i\in[d]}$ denote the identity basis, put $a=r+1$, and define $\mathcal{D}_0$ by $(x_i,y_i)=(e_i,0)$ for $i\in[r]$ and $(x_{r+1},y_{r+1})=(0,1)$. Then $\mathcal{D}_0$ is Δ-fixed with $\Delta=1$ and contact vectors $\{\pm e_i\}_{i\in[r]}$, resulting in $\lambda=2$. Now take $\sigma^2=1/(r\epsilon)$ so that $\mathsf{SNR}=\epsilon$, and draw $\mathcal{D}\sim\operatorname{PFM}(\mathcal{D}_0,\sigma^2,b)$. By our assumptions on ϵ and α, we may select $\beta\in(\frac{\epsilon\alpha}{2},\min\{\alpha,1-\alpha\})$, $p=\lfloor\beta d\rfloor$, and let Π_p denote orthogonal projection onto $\operatorname{span}\{e_i\}_{i=r+1}^{r+p}$. We claim that, unless b is superpolynomial in d, then for sufficiently large d, it holds with probability $\geq 1/2$ that Π_p is feasible in $\operatorname{sqz}(\mathcal{D},\Delta)$. Since $\operatorname{tr}\Pi_p=p<r=\operatorname{tr}\Pi$, we then have $\Pi\notin\arg\operatorname{sqz}(\mathcal{D},\Delta)$.

To proceed, we estimate

$$\Delta_p^2:=\min_{\substack{i\in[r]\\ s,t\in[b]}}\Big\|\Pi_p\Big((x_i+g_{is})-(x_{r+1}+g_{r+1,t})\Big)\Big\|^2=\min_{\substack{i\in[r]\\ s,t\in[b]}}\|\Pi_p(g_{is}-g_{r+1,t})\|^2.$$

First, for every $i\in[r]$ and $s,t\in[b]$, expanding the square gives

$$\begin{aligned}\|\Pi_p(g_{is}-g_{r+1,t})\|^2&=\|\Pi_p g_{is}\|^2-2\cdot\|\Pi_p g_{r+1,t}\|\cdot\langle\Pi_p g_{is},\tfrac{\Pi_p g_{r+1,t}}{\|\Pi_p g_{r+1,t}\|}\rangle+\|\Pi_p g_{r+1,t}\|^2\\&\geq 2\min_{\substack{i\in[a]\\ s\in[b]}}\|\Pi_p g_{is}\|^2-2\Big(\max_{t\in[b]}\|\Pi_p g_{r+1,t}\|\Big)\Big(\max_{\substack{i\in[r]\\ s,t\in[b]}}|\langle\Pi_p g_{is},\tfrac{\Pi_p g_{r+1,t}}{\|\Pi_p g_{r+1,t}\|}\rangle|\Big).\end{aligned}$$

Next, recall that if Z is standard Gaussian and Q is χ^2-distributed with q degrees of freedom, then

$$\Pr(|Z|\geq\xi)\leq 2e^{-\xi^2/2},\qquad\Pr(|Q-q|\geq\xi)\leq 2e^{-c\min\{\xi,\xi^2/q\}},\qquad\xi\geq 0\tag{9}$$

for some universal constant $c>0$; in particular, these estimates follow from the Chernoff bound and Hanson–Wright inequality [32], respectively. We apply these bounds to obtain the estimate

$$\Delta_p^2\geq\sigma^2\cdot\Big(2(p-\xi_1)-2(p+\xi_1)^{1/2}\cdot\xi_2\Big)$$

with probability $\geq 1-2abe^{-c\min\{\xi_1,\xi_1^2/p\}}-2rb^2e^{-\xi_2^2/2}$. If we select $\xi_1=p^{3/4}$ and $\xi_2=p^{1/4}$, then we may conclude $\Delta_p^2\geq 2\beta/(\epsilon\alpha)-o_{d\to\infty}(1)$ with probability $\geq 1/2$ unless b is superpolynomial in d. Since $\beta>\epsilon\alpha/2$, we have $\Delta_p^2>1=\Delta^2$ for large d, which implies Π_p is feasible in $\operatorname{sqz}(\mathcal{D},\Delta)$. □

Proof of Corollary 9. Set

$$b_1 := \left\lceil \max\left\{\frac{2}{c_2}(d-r), c_3\right\}\right\rceil, \qquad \mathsf{SNR}_1 := 3c_1\Big(4+\frac{c_2}{2}\Big)\sqrt{\log b_1}, \qquad \sigma_1^2 := \frac{\lambda}{r\cdot \mathsf{SNR}_1}.$$

If $\sigma \leq \sigma_1$, we may take $b = b_1$ by Theorem 7. Otherwise, set $p := (\sigma_1/\sigma)^{d-r}$, and select b large enough so that for independent Bernoulli random variables $\{B_{it}\}_{i\in[a],t\in[b]}$ with mean p, it holds with high probability that $\sum_{t\in[b]} B_{it} \geq b_1$ for every $i \in [a]$. To see why this suffices, observe that there exists a distribution $\mathcal{F}$ such that

$$\mathcal{N}(0,\sigma^2(I-\Pi)) = p\cdot\mathcal{N}(0,\sigma_1^2(I-\Pi)) + (1-p)\cdot\mathcal{F}.$$

Let $\mathcal{D}_1$ denote the data points in $\mathcal{D}$ corresponding to the $\mathcal{N}(0,\sigma_1^2(I-\Pi))$ component. In the high-probability event that $\sum_{t\in[b]} B_{it} \geq b_1$ for every $i\in[a]$, Theorem 7 gives that $\arg\mathrm{sqz}(\mathcal{D}_1,\Delta) = \{\Pi\}$ with high probability, which is also feasible in $\mathrm{sqz}(\mathcal{D},\Delta)$. Since $\mathrm{sqz}(\mathcal{D}_1,\Delta)$ is a relaxation of $\mathrm{sqz}(\mathcal{D},\Delta)$, we may conclude that $\arg\mathrm{sqz}(\mathcal{D},\Delta) = \{\Pi\}$ in the same event. □

Proof of Theorem 7. By Lemma 10(i), $\arg\mathrm{sqz}(\mathcal{D}_0,\Delta) = \{\Pi\}$, which is trivially feasible in $\mathrm{sqz}(\mathcal{D},\Delta)$. We will modify a dual certificate for $\mathrm{sqz}(\mathcal{D}_0,\Delta)$ to produce a point (γ, Y) in $\mathrm{dual}(\mathcal{D},\Delta)$ of the same value. By the proof of Theorem 15, $\mathrm{sqz}(\mathcal{D}_0,\Delta)$ enjoys a dual certificate $(\tilde\gamma,\tilde Y)$ of the form

$$\tilde\gamma(z) := \left\{\begin{array}{cl} 1/\lambda & \text{if } z\in\mathcal{Z}_0 \\ 0 & \text{if } z\in\mathcal{Z}(\mathcal{D}_0)\setminus\mathcal{Z}_0\end{array}\right\}, \qquad \tilde Y := \frac{1}{\lambda}\sum_{z\in\mathcal{Z}_0} zz^\top - \Pi,$$

Our choice of (γ, Y) will take $Y = \tilde Y$, but selecting an appropriate γ will be more delicate. Denote $A_\gamma := \sum_{z\in\mathcal{Z}(\mathcal{D})}\gamma(z)zz^\top - Y$ and $\Pi_\perp := I - \Pi$. We will select $\gamma\geq 0$ such that

$$\|\Pi A_\gamma \Pi\|_{2\to2}\leq 1, \quad \|\Pi_\perp A_\gamma\Pi_\perp\|_{2\to2}\leq 1, \quad \Pi A_\gamma\Pi_\perp = 0, \qquad \sum_{z\in\mathcal{Z}(\mathcal{D})}\gamma(z) = \sum_{z\in\mathcal{Z}(\mathcal{D}_0)}\tilde\gamma(z).$$

The first three above together imply $A_\gamma \preceq I$, thereby ensuring (γ, Y) is feasible in $\mathrm{dual}(\mathcal{D},\Delta)$, while the final condition ensures that the value of (γ, Y) in $\mathrm{dual}(\mathcal{D},\Delta)$ equals the value of $(\tilde\gamma,\tilde Y)$ in $\mathrm{dual}(\mathcal{D}_0,\Delta)$, as desired.

We first claim that with high probability, it suffices to have the following: For every $i,j\in[a]$ such that $x_i - x_j\in\mathcal{Z}_0$, there exists $v_{ij}\in\mathbb{R}^b$ such that

$$v_{ij}\geq 0, \qquad 1^\top v_{ij} = \frac{1}{\lambda}, \qquad \max v_{ij}\leq \frac{c_1\sqrt{\log b}}{b\lambda}, \qquad G_{ij}v_{ij} = 0, \tag{10}$$

where G_{ij} is the $d\times b$ matrix whose tth column is $g_{it} - g_{jt}$. To see this, consider γ defined by

$$\gamma\Big((x_i+g_s) - (x_j+g_t)\Big) = \left\{\begin{array}{cl} (v_{ij})_t & \text{if } x_i - x_j\in\mathcal{Z}_0 \text{ and } s = t \\ 0 & \text{otherwise.}\end{array}\right.$$

Then $\gamma\geq 0$ is immediate, while $\|\Pi A_\gamma\Pi\|_{2\to2}\leq 1$ follows from

$$\Pi A_\gamma \Pi = \sum_{\substack{i,j\in[a] \\ x_j - x_j\in\mathcal{Z}_0}}\sum_{t\in[b]}(v_{ij})_t(x_i-x_j)(x_i-x_j)^\top - \Pi Y\Pi = \Pi\Big(\frac{1}{\lambda}X - \tilde Y\Big)\Pi = \Pi.$$

Next, (5) implies $\Pi_\perp Y = 0$, and so

$$\begin{aligned}\|\Pi_\perp A_\gamma \Pi_\perp\|_{2\to 2} &= \Bigg\| \sum_{\substack{i,j\in[a]\\ x_j - x_j \in \mathcal{Z}_0}} \sum_{t\in[b]} (v_{ij})_t (g_{it} - g_{jt})(g_{it} - g_{jt})^\top \Bigg\|_{2\to 2} \\ &\leq \frac{c_1\sqrt{\log b}}{b\lambda} \sum_{\substack{i,j\in[a]\\ x_j - x_j \in \mathcal{Z}_0}} \Bigg\| \sum_{t\in[b]} (g_{it} - g_{jt})(g_{it} - g_{jt})^\top \Bigg\|_{2\to 2} \\ &\leq \frac{c_1\sqrt{\log b}}{b\lambda} \cdot r \cdot 2\sigma^2 \cdot \Big(\sqrt{d-r} + 2\sqrt{b}\Big)^2 \leq \frac{3c_1\sqrt{\log b}}{\mathsf{SNR}} \cdot \Big(\frac{d-r}{b} + 4\Big) \leq 1,\end{aligned}$$

where the last line applies Corollary 5.35 in [37] with a union bound, the inequality $(p+q)^2 \leq 3(p^2+q^2)$, and finally our assumptions on b and SNR; in particular, the first inequality in this last line is valid in an event $\mathcal{E}_1$ of probability $\geq 1 - 2e^{-b/2}$. Finally, we also have

$$\Pi A_\gamma \Pi_\perp = \sum_{\substack{i,j\in[a]\\ x_j - x_j \in \mathcal{Z}_0}} \sum_{t\in[b]} (v_{ij})_t (x_i - x_j)(g_{it} - g_{jt})^\top = \sum_{\substack{i,j\in[a]\\ x_j - x_j \in \mathcal{Z}_0}} (x_i - x_j)(G_{ij} v_{ij})^\top = 0,$$

and

$$\sum_{z\in\mathcal{Z}(\mathcal{D})} \gamma(z) = \sum_{\substack{i,j\in[a]\\ x_j - x_j \in \mathcal{Z}_0}} \sum_{t\in[b]} (v_{ij})_t = \frac{|\mathcal{Z}_0|}{\lambda} = \sum_{z\in\mathcal{Z}(\mathcal{D}_0)} \tilde{\gamma}(z).$$

It remains to find v_{ij}'s that satisfy (10). Equivalently, we must show that with high probability, it holds that for every $i,j \in [a]$ such that $x_i - x_j \in \mathcal{Z}_0$, the random subspace $\operatorname{Null}(G_{ij})$ nontrivially intersects the cone $\mathcal{C}_b$ defined in Lemma 6. Importantly, each $\operatorname{Null}(G_{ij})$ has the same distribution as $\operatorname{Null}(G)$, where G is $(d-r)\times b$ with independent standard Gaussian entries, meaning $\operatorname{Null}(G_{ij})$ is drawn uniformly from the Grassmannian of $(b-d+r)$-dimensional subspaces of $\mathbb{R}^b$. Then Theorem 7.1 in [1] gives

$$d - r \leq \delta(\mathcal{C}_b) - \xi \qquad \Longrightarrow \qquad \Pr\Big(\mathcal{C}_b \cap \operatorname{Null}(G_{ij}) = \{0\}\Big) \leq 4\exp\left(\frac{-\xi^2/8}{\delta(\mathcal{C}_b)+\xi}\right)$$

whenever $\xi \geq 0$. Recalling Lemma 6 and selecting $\xi = c_2 b/2$, then since $d - r \leq c_2 b/2$ by assumption, we obtain nontrivial intersection between $\mathcal{C}_b$ and each $\operatorname{Null}(G_{ij})$ in an event $\mathcal{E}_2$ of probability $\geq 1 - 4|\mathcal{Z}_0|e^{-c_2 b/48}$. The result then follows by a union bound between $\mathcal{E}_1$ and $\mathcal{E}_2$. □

Proof of Lemma 6. First, $\mathcal{C}_n$ is contained in the self-dual nonnegative orthant $\mathbb{R}^n_+$. As such, Propositions 3.1 and 3.2 in [1] together give

$$\delta(\mathcal{C}_n) \leq \delta(\mathbb{R}^n_+) = \frac{1}{2}n.$$

Next, given any bounded set $T \subseteq \mathbb{R}^n$, let $N(T,\epsilon)$ denote the size of the largest ϵ-packing, that is, the largest $S \subseteq T$ such that

$$\min_{\substack{x,y\in S\\ x\neq y}} \|x - y\| \geq \epsilon.$$

By Sudakov minoration (see the proof of Theorem 7.4.1 in [36], for example), we may bound statistical dimension in terms of packing numbers:

$$\delta(\mathcal{C}_n) \geq \Big(\mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,I)} \sup_{x\in\mathcal{C}_n\cap\mathbb{S}^{n-1}} \langle x, g\rangle \Big)^2 \geq c \cdot \sup_{\epsilon>0} \epsilon^2 \log N(\mathcal{C}_n \cap \mathbb{S}^{n-1}, \epsilon), \tag{11}$$

where $c > 0$ is some universal constant (one may take $c = 1 - e^{-1}$, for example). As such, for the remaining lower bound, it suffices to estimate packing numbers of $\mathcal{C}_n \cap \mathbb{S}^{n-1}$.

To this end, we make a general observation: Fix a measurable set $T \subseteq \mathbb{S}^{n-1}$ of normalized surface area $s(T) := \operatorname{area}(T)/\operatorname{area}(\mathbb{S}^{n-1})$, let P denote a largest ϵ-packing of $\mathbb{S}^{n-1}$, and find a rotation $Q_0 \in \mathrm{SO}(n)$ that maximizes the cardinality of $Q_0 P \cap T$. If we draw Q uniformly from $\mathrm{SO}(n)$, then the linearity of expectation gives

$$N(T, \epsilon) \geq |Q_0 P \cap T| \geq \mathbb{E}|QP \cap T| = \mathbb{E} \sum_{x \in P} \mathbf{1}_{\{Qx \in T\}} = \sum_{x \in P} \Pr(Qx \in T) = s(T) \cdot N(\mathbb{S}^{n-1}, \epsilon). \quad (12)$$

A volume comparison argument gives the following estimate:

$$N(\mathbb{S}^{n-1}, 2\sin\tfrac{\theta}{2}) \geq (1 + o_{n\to\infty}(1)) \cdot \sqrt{2\pi n} \cdot \frac{\cos\theta}{\sin^{n-1}\theta} \qquad \forall \theta \in (0, \tfrac{\pi}{2}). \quad (13)$$

(See [21] for a recent improvement to this estimate, and references therein for historical literature related to (13), which will suffice for our purposes.) With this, one obtains a lower bound on $N(T, \epsilon)$ by computing $s(T)$, which equals the probability that $g \sim \mathcal{N}(0, I)$ resides in the positively homogeneous set $\bigcup_{r>0} rT$ generated by T.

In our special case of $T = \mathcal{C}_n \cap \mathbb{S}^{n-1}$, we condition on the event $\{g \geq 0\}$ to obtain

$$s(\mathcal{C}_n \cap \mathbb{S}^{n-1}) = \Pr(g \in \mathcal{C}_n) = 2^{-n} \cdot \Pr\Big(\max h \leq \frac{c_1\sqrt{\log n}}{n} \cdot 1^\top h \Big),$$

where there coordinates of h are independent with standard half-normal distribution. Next, for every choice of $\alpha > 0$, the union bound gives

$$\begin{aligned}\Pr\Big(\max h \leq \frac{c_1\sqrt{\log n}}{n} \cdot 1^\top h \Big) &\geq \Pr\Big(\max h < \alpha\sqrt{\log n} < \frac{c_1\sqrt{\log n}}{n} \cdot 1^\top h \Big) \\ &\geq 1 - \Pr\Big(\max h \geq \alpha\sqrt{\log n} \Big) - \Pr\Big(\frac{c_1}{n} 1^\top h \leq \alpha \Big).\end{aligned}$$

We apply another union bound with (9) to estimate the first term:

$$\Pr\Big(\max h \geq \alpha\sqrt{\log n} \Big) \leq n \cdot \Pr(|Z| \geq \alpha\sqrt{\log n}) \leq \frac{2}{n^{\alpha^2/2-1}}.$$

To estimate the second term, we use the fact that each coordinate h_i of h has mean $\sqrt{2/\pi}$ and variance $1 - 2/\pi$ and apply Chebyshev's inequality:

$$\Pr\Big(\frac{c_1}{n} 1^\top h \leq \alpha \Big) \leq \Pr\Bigg(\bigg| \frac{1}{n} \sum_{i \in [n]} h_i - \sqrt{\frac{2}{\pi}} \bigg| \geq \sqrt{\frac{2}{\pi}} - \frac{\alpha}{c_1} \Bigg) \leq \frac{1 - 2/\pi}{(\sqrt{2/\pi} - \alpha/c_1)^2}.$$

Combining these estimates, we may select $\alpha = 4$ and $c_1 = 50$ (say) to get $s(\mathcal{C}_n \cap \mathbb{S}^{n-1}) \geq \frac{1}{4} \cdot 2^{-n}$ for every $n \geq 2$. Then taking $\theta = \frac{\pi}{12}$ and combining with (12) and (13) gives

$$\log N(\mathcal{C}_n \cap \mathbb{S}^{n-1}, 2\sin\tfrac{\pi}{24}) \geq 0.65n + \tfrac{1}{2}\log n - 1.86 + o_{n\to\infty}(1).$$

By (11), we are done. □

4 Numerical experiments

4.1 Implementation variants

Before describing our numerical experiments, we first discuss a few different implementations of SqueezeFit that allow for scalability and robustness to outliers.

Hinge loss. SqueezeFit is not feasible if Δ is larger than the minimum distance between two points with different labels. In order to make SqueezeFit robust to outliers, we replace the Δ constraints in $\mathrm{sqz}(\mathcal{D},\Delta)$ with a hinge-loss penalization in the objective:

$$\text{minimize} \quad \operatorname{tr} M + \lambda \sum_{z\in\mathcal{Z}(\mathcal{D})} \Big(\Delta^2 - z^\top M z\Big)_+ \quad \text{subject to} \quad 0 \preceq M \preceq I \qquad (\mathrm{sqz}_\lambda(\mathcal{D},\Delta))$$

Here, $a_+ := \max\{0,a\}$. Importantly, $\mathrm{sqz}_\lambda(\mathcal{D},\Delta)$ is feasible regardless of Δ, but this comes at the price of an additional hyperparameter $\lambda > 0$.

Relaxing Δ constraints. Suppose $\mathcal{D} = \{(x_i, y_i)\}_{i\in[n]}$ is comprised of n data points that are balanced over k labels. Then $\mathrm{sqz}(\mathcal{D},\Delta)$ is a semidefinite program with d^2 variables and $O_k(n^2)$ constraints. When implemented directly, we find this program to be too slow once $n \geq 200$, even when d is small. However, Lemma 17 indicates that typically, most of these Δ constraints are not tight, and so to accommodate larger values of n, we relax many of these constraints. Fix $s \geq 1$, let $S(i,\ell)$ denote the indices of the s nearest neighbors to x_i with label ℓ, and put

$$\mathcal{Z}_s(\mathcal{D}) := \bigcup_{i\in[n]} \bigcup_{\substack{\ell\in[k]\\ \ell\neq y_i}} \Big\{x_i - x_j : j \in S(i,\ell)\Big\}.$$

Then replacing $\mathcal{Z}(\mathcal{D})$ in $\mathrm{sqz}(\mathcal{D},\Delta)$ with $\mathcal{Z}_s(\mathcal{D})$ results in a relaxation with only $O(sn)$ constraints. In practice, we obtain $\mathcal{Z}_s(\mathcal{D})$ in $O_d(n \log n)$ time using a k-d tree. Overall, passing to $\mathcal{Z}_s(\mathcal{D})$ results in the following variants:

$$\text{minimize} \quad \operatorname{tr} M \quad \text{subject to} \quad z^\top M z \geq \Delta^2 \quad \forall z \in \mathcal{Z}_s(\mathcal{D}), \quad 0 \preceq M \preceq I \qquad (\mathrm{sqz}^s(\mathcal{D},\Delta))$$

$$\text{minimize} \quad \operatorname{tr} M + \lambda \sum_{z\in\mathcal{Z}_s(\mathcal{D})} \Big(\Delta^2 - z^\top M z\Big)_+ \quad \text{subject to} \quad 0 \preceq M \preceq I \qquad (\mathrm{sqz}^s_\lambda(\mathcal{D},\Delta))$$

As one might expect, we observe that the optimizers of these variants are close approximations to optimizers of the original programs, even for moderate values of s, and the approximation is especially good when $\mathcal{D}$ exhibits clustering structure.

Relaxing the identity constraint. For every $\mathcal{D}$, there exists $\delta = \delta(\mathcal{D}) > 0$ such that for every $\Delta \leq \delta$, it holds that every $M \in \arg\mathrm{sqz}(\mathcal{D},\Delta)$ satisfies $\operatorname{tr} M < 1$, meaning the constraint $M \preceq I$ is not tight. As such, we can afford to relax the identity constraint when Δ is small. In the absence of the identity constraint, then up to scaling, we may equivalently put $\Delta = 1$. For this reason, we define the variant

$$\text{minimize} \quad \operatorname{tr} M \quad \text{subject to} \quad z^\top M z \geq 1 \quad \forall z \in \mathcal{Z}(\mathcal{D}), \quad M \succeq 0 \qquad (\mathrm{sqz}(\mathcal{D},0^+))$$

We similarly define $\mathrm{sqz}_\lambda(\mathcal{D},0^+)$, etc. We observe that solving this relaxation is considerably faster.

4.2 Handwritten digits

In this subsection, we use SqueezeFit to perform compressive classification on the MNIST database of handwritten digits [23]. Specifically, we focus on binary classification between 4s and 9s, since

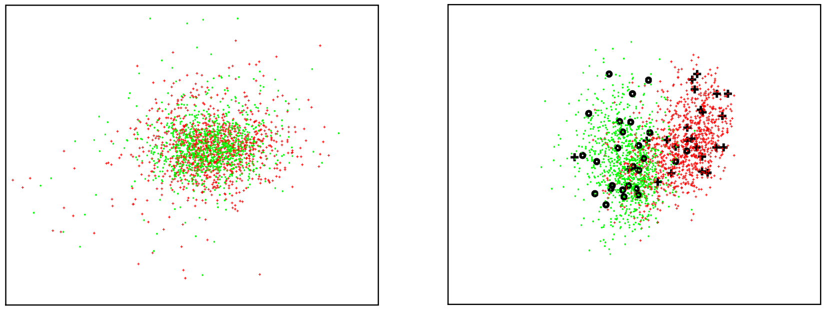

Figure 3: Take low-resolution versions of MNIST digits (4s and 9s) to obtain images in $[0,1]^{10\times 10}$. The two leading principal components of this data are displayed on the left. Letting $\mathcal{D}$ denote 50 of these low-resolution images, we run a variant of SqueezeFit to obtain M of rank 5. Letting Π denote orthogonal projection onto the span of the two leading eigenvectors of M, the right-hand plot illustrates Πx for 2000 randomly selected x's from the low-resolution 4s and 9s. The 50 points from $\mathcal{D}$ are indicated with '+' and 'o'. Impressively, SqueezeFit identifies components that keep all of the 4s and 9s separated, despite only seeing a small sample.

these digits are easily confused. There are 11,971 of such digits in the training set and 1,991 in the test set. In order to apply SqueezeFit, we first decrease the dimensionality of the space by forming low-resolution versions of these digits in $[0,1]^{10\times 10}$. Next, we select n points $\mathcal{D}$ at random from the training set and compute $M \in \arg\operatorname{sqz}_\lambda^s(\mathcal{D}, 0^+)$ with $\lambda = 1$ and $s = 5$. We then apply $M^{1/2}$ to the entire training set. Figure 3 illustrates this SqueezeFit compression with $n = 50$. After compression, we apply K-nearest neighbor classification on the test set. Figure 4 illustrates this classification in the case of $n = 800$ and $K = 15$. Table 1 compares the misclassification rates for K-nearest neighbor classification after compression with PCA, LDA, NuMax, and SqueezeFit. This comparison indicates that SqueezeFit is competitive with NuMax at finding low-dimensional components that are amenable to classification.

4.3 Hyperspectral imagery

The Indian Pines hyperspectral dataset [19] consists of a $145 \times 145 \times 200$ data cube, representing a 145×145 overhead scene of farm land with 200 different spectral reflectance bands ranging from 0.4 to 2.5 micrometers. Each of the 145^2 pixels in this scene is labeled by a member of $\{0, 1, \dots, 16\}$; labels 1 through 16 either correspond to some sort of crop or some other material, whereas the label 0 means the pixel is not labeled. Since real-world hyperspectral data collection is slow, we wish to classify the contents of a pixel in a hyperspectral image from as few spectral measurements as possible. For simplicity, we consider the binary classification task of distinguishing crops from non-crops. In order to evaluate per-pixel compressive classification, we split the Indian Pines pixels with nonzero labels into 70% training data and 30% testing data. Much like we did for MNIST digits above, we downsampled each 200-dimensional feature vector into a point in $\mathbb{R}^{100}$. We then applied PCA, LDA, NuMax, and SqueezeFit to a subset of the training set in order to compute a compression operator, and we then applied this operator to the entire training set before performing K-nearest neighbor classification on the test set. The results are summarized in Table 2. Much like Table 1, this comparison indicates that SqueezeFit is well suited for finding low-dimensional components for classification.

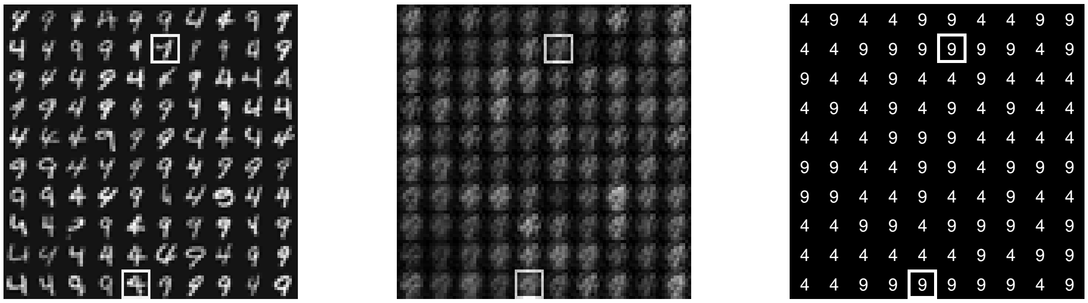

Figure 4: **(left)** Low-resolution versions of 4s and 9s from the MNIST test set. **(middle)** Given 800 random members of the MNIST training set, run a variant of SqueezeFit to obtain M of rank 11, and then apply $M^{1/2}$ to the digits in the left panel. **(right)** Apply K-nearest neighbor classification with $K = 15$ to the compressed digits in the middle panel to predict labels. Squares indicate misclassified images.

	Id	**PCA**		**LDA**		**NuMax**		**SqueezeFit**	
n	0	11,791	11,791	800	11,791	50	800	50	800
r	100	5	11	1	1	6	14	5	11
$K = 1$	2.15	16.62	5.47	7.78	6.07	7.03	2.96	5.97	4.01
$K = 5$	1.95	12.55	4.26	5.32	4.62	5.47	2.31	5.47	3.61
$K = 15$	2.26	12.00	4.11	5.32	3.87	5.67	2.51	5.27	3.87

Table 1: Percentage of misclassified MNIST digits (4s and 9s) from K-nearest neighbor classification after compressing with either PCA, LDA, NuMax, or SqueezeFit. For each column, n points from the training set were used to find a compression operator of rank r. (The Id column takes identity to be the "compression" operator.) This compression operator was then applied to the entire training set (of size 11,971) before running K-nearest neighbor classification on the test set (of size 1,991). For each $n \in \{50, 800\}$, the corresponding columns computed a compression operator based on the same random sample of the training set.

5 Discussion

In this paper, we introduced projection factor recovery as an idealization of the compressive classification problem, we proposed SqueezeFit as a semidefinite programming approach to this problem, and we provided theoretical guarantees for SqueezeFit in the context of projection factor recovery, as well as numerical experiments that compare SqueezeFit to alternative methods for compressive classification. Through this investigation, the authors encountered a trove interesting research questions and opportunities for future work, which we discuss below.

First, under what conditions is projection factor recovery possible, both computationally and information theoretically? In this paper, we focused on the case where SqueezeFit is well suited to perform projection factor recovery. In particular, we established that the SNR threshold in Theorem 7 is tight up to logarithmic factors, but is the $\sqrt{\log b}$ factor necessary? Importantly, the projection factor model allows for exact projection factor recovery when b is not too small (e.g., when $b > d - r$). However, when $b = 1$, exact projection factor recovery is no longer possible,

	Id	PCA		LDA		NuMax		SqueezeFit	
n	0	7,175	7,175	300	7,175	100	300	100	300
r	100	3	5	1	1	4	6	3	5
$K=1$	1.39	4.81	2.76	4.81	3.96	3.83	2.04	2.73	2.11
$K=5$	1.69	4.35	2.79	3.12	3.02	3.41	2.08	2.83	2.27
$K=15$	2.14	4.81	3.38	2.89	2.86	3.51	2.50	2.96	2.83

Table 2: Percentage of misclassified pixels in the Indian Pines test set from K-nearest neighbor classification after compressing with either PCA, LDA, NuMax, or SqueezeFit. For each column, n points from the training set were used to find a compression operator of rank r. (The Id column takes identity to be the "compression" operator.) This compression operator was then applied to the entire training set (of size 7,175) before running K-nearest neighbor classification on the test set (of size 3,074). For each $n \in \{100, 300\}$, the corresponding columns computed a compression operator based on the same random sample of the training set.

although we observe approximate recovery in Figure 1. Under what conditions does SqueezeFit give approximate recovery in this model?

While SqueezeFit has proven to be particularly amenable to theoretical investigation, our numerical experiments encountered a barrier to running the semidefinite program on large data. For instance, it takes about 50 minutes to run SqueezeFit on a 100-dimensional dataset comprised of 800 data points (on a standard Macbook Air 2013). This limitation forced us to randomly sample the training sets before running SqueezeFit (mimicking [29]), but presumably, one can devise better sampling techniques based on some choice of leverage scores that account for the full geometry of the training set. Without such sampling, then in order to handle datasets with more points in higher dimensions (e.g., the full MNIST training set [23]), we require a different approach to solving (1). For example, one might consider a reformulation of (1) that optimizes over the $O(rd)$-dimensional Grassmannian of r-dimensional subspaces of $\mathbb{R}^d$ instead of the $\Omega(d^2)$-dimensional cone of positive semidefinite $d \times d$ matrices. While such a formulation will be non-convex, there is a growing body of work that provides performance guarantees for such optimization problems [8, 24, 28, 5]. In fact, [35] proposes such a non-convex formulation of LMNN, and it performs well in practice.

Finally, there is room to further improve SqueezeFit for more effective compressive classification. In practice, one will encounter application-specific design constraints on the sensing operator, and it would be interesting to incorporate these constraints into the SqueezeFit program. For example, if the sensor A is required to be a linear filter, then $M = A^\top A$ is diagonalized by the discrete Fourier transform, and so SqueezeFit reduces to a linear program. This constrained formulation enjoys runtime speedups over the original semidefinite program (see [13]), and presumably, one may obtain refined performance guarantees. Also, the numerical experiments in this paper focused on k-nearest neighbor classifiers in order to isolate the performance of dimensionality reduction alternatives. However, the best known algorithms for image classification use convolutional neural networks (see [22], for example), and so it would be interesting to impose relevant convolution-friendly constraints in the SqueezeFit program to make use of this performance (see [25] for related work).

Acknowledgments

The authors thank the anonymous reviewers for various suggestions that substantially improved our paper. CM and DGM were partially supported by the Center for Surveillance Research at the Ohio State University (an NSF industry/university cooperative research center). DGM was partially supported by AFOSR FA9550-18-1-0107, NSF DMS 1829955, and the Simons Institute of the Theory of Computing. SV was partially supported by EOARD FA9550-18-1-7007 and by the Simons Algorithms and Geometry (A&G) Think Tank. The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

References

[1] D. Amelunxen, M. Lotz, M. B. McCoy, J. A. Tropp, Living on the edge: Phase transitions in convex programs with random data, Inform. Inference 3 (2014) 224–294.

[2] A. S. Bandeira, A note on probably certifiably correct algorithms, C. R. Math. 354 (2016) 329–333.

[3] A. S. Bandeira, D. G. Mixon, B. Recht, Compressive classification and the rare eclipse problem, In: Compressed Sensing and its Applications, pp. 197–220. Birkhäuser, Cham, 2017.

[4] P. T. Boufounos, R. G. Baraniuk, 1-bit compressive sensing, CISS 2008, 16–21.

[5] N. Boumal, V. Voroninski, A. S. Bandeira, Deterministic guarantees for Burer-Monteiro factorizations of smooth semidefinite programs, arXiv:1804.02008

[6] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge, 2004.

[7] E. J. Candès, Y. C. Eldar, T. Strohmer, V. Voroninski, Phase retrieval via matrix completion, SIAM Rev. 57 (2015) 225–251.

[8] E. J. Candès, X. Li, M. Soltanolkotabi, Phase retrieval via Wirtinger flow: Theory and algorithms, IEEE Trans. Inform. Theory 61 (2015) 1985–2007.

[9] E. J. Candès, J. K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Comm. Pure Appl. Math. 59 (2006) 1207–1223.

[10] E. J. Candès, T. Strohmer, V. Voroninski, Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming, Comm. Pure Appl. Math. 66 (2013) 1241–1274.

[11] M. A. Davenport, M. F. Duarte, M. B. Wakin, J. N. Laska, D. Takhar, K. F. Kelly, R. G. Baraniuk, The smashed filter for compressive classification and target recognition, Proc. SPIE, Computational Imaging 6498 (2007) 64980H.

[12] D. L. Donoho, Compressed sensing, IEEE Trans. Inform. Theory, 52 (2006) 1289–1306.

[13] B. Dumitrascu, S. Villar, D. G. Mixon, B. E. Engelhardt, Optimal marker gene selection for cell type discrimination in single cell analyses bioRxiv:599654

[14] S. Foucart, H. Rauhut, A mathematical introduction to compressive sensing, Basel: Birkhäuser, 2013.

[15] J. H. Friedman, J. L. Bentley, R. A. Finkel, An algorithm for finding best matches in logarithmic time, ACM Trans. Math. Software 3 (1976) 209–226.

[16] M. A. Goberna, M. A. Lopez, Linear semi-infinite programming theory: An updated survey, Eur. J. Oper. Res. 143 (2002) 390–405.

[17] Y. Han, Tight bound for matching, J. Comb. Optim. (2012) 23:322–330.

[18] C. Hegde, A. C. Sankaranarayanan, W. Yin, R. G. Baraniuk, NuMax: A convex approach for learning near-isometric linear embeddings, IEEE Trans. Signal Process. 63 (2015) 6109–6121.

[19] Hyperspectral Remote Sensing Scenes, `ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes`

[20] T. Iguchi, D. G. Mixon, J. Peterson, S. Villar, Probably certifiably correct k-means clustering, Math. Program. 165 (2017) 605–642.

[21] M. Jenssen, F. Joos, W. Perkins, On kissing numbers and spherical codes in high dimensions, arXiv:1803.02702

[22] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012, 1097–1105

[23] Y. LeCun, C. Cortes, C. J. Burges, MNIST handwritten digit database, `yann.lecun.com/exdb/mnist`

[24] X. Li, S. Ling, T. Strohmer, K. Wei, Rapid, robust, and reliable blind deconvolution via nonconvex optimization, Appl. Comput. Harmon. Anal. (2018).

[25] Y. Li, D. Pinckney, T. Strohmer, Compressive Deep Learning, in preparation.

[26] R. Litiu, D. I. Kountanis, Closest Pair for Two Separated Sets of Points, Congressus Numerantium (1997) 97–112.

[27] M. Lustig, D. Donoho, J. M. Pauly, Sparse MRI: The application of compressed sensing for rapid MR imaging, Magn. Reson. Med. 58 (2007) 1182–1195.

[28] T. Maunu, T. Zhang, G. Lerman, A well-tempered landscape for non-convex robust subspace recovery, arXiv:1706.03896

[29] D. G. Mixon, S. Villar, Monte Carlo approximation certificates for k-means clustering, arXiv:1710.00956

[30] H. Reboredo, F. Renna, R. Calderbank, M. R. D. Rodrigues, Bounds on the number of measurements for reliable compressive classification, IEEE Trans. Signal Process. 64 (2016) 5778–5793.

[31] D. M. Rosen, L. Carlone, A. S. Bandeira, J. J. Leonard, SE-Sync: A certifiably correct algorithm for synchronization over the special Euclidean group, arXiv:1612.07386

[32] M. Rudelson, R. Vershynin, Hanson-Wright inequality and sub-gaussian concentration, Electronic Communications in Probability 18 (2013).

[33] P. Sidorenko, O. Kfir, Y. Shechtman, A. Fleischer, Y. C. Eldar, M. Segev, O. Cohen, Sparsity-based super-resolved coherent diffraction imaging of one-dimensional objects, Nature Comm. 6 (2015) 8209.

[34] D. Silver, et al., Mastering the game of Go with deep neural networks and tree search, Nature 529 (2016) 484.

[35] L. Torresani, K.-C. Lee, Large margin component analysis, NIPS 2007, 1385–1392.

[36] R. Vershynin, High-dimensional probability: An introduction with applications in data science, Cambridge, 2018.

[37] R. Vershynin, Introduction to the non-asymptotic analysis of random matrices, In: Compressed Sensing, Theory and Applications, Y. Eldar, G. Kutyniok (eds.), Cambridge, 2012, 210–268.

[38] V. Q. Vu, J. Cho, J. Lei, K. Rohe, Fantope projection and selection: A near-optimal convex relaxation of sparse PCA, NIPS 2013, 2670–2678.

[39] K. Q. Weinberger, L. K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (2009) 207–244.

[40] Z. Yang, K. Balasubramanian, H. Liu, On Stein's Identity and Near-Optimal Estimation in High-dimensional Index Models, arXiv:1709.08795