INFORMATION-THEORETIC CRITERIA FOR THE DESIGN OF COMPRESSIVE SUBSPACE CLASSIFIERS

Matthew Nokleby[†]*, Miguel Rodrigues*[⋆]*, and Robert Calderbank*[†]

[†]Duke University, Durham, NC, emails: {matthew.nokleby, robert.calderbank}@duke.edu
[⋆]University College London, UK, email: m.rodrigues@ucl.ac.uk

ABSTRACT

Using Shannon theory, we derive fundamental, asymptotic limits on the classification of low-dimensional subspaces from compressive measurements. We identify a syntactic equivalence between the classification of subspaces and the communication of codewords over non-coherent, multiple-antenna channels, from which we derive sharp bounds on the number of classes that can be discriminated with low misclassification probability as a function of the signal dimensionality and the signal-to-noise ratio. While the bounds are asymptotic in the limit of high dimension, they provide intuition for classifier design at finite dimension. We validate this intuition via an application to face recognition.

1. INTRODUCTION

Compressive sensing offers the means to simultaneously sense and compress a high-dimensional signal that belongs to a low-dimensional manifold. The paradigmatic result of compressed sensing is that one can recover, with overwhelming probability, a signal that is k-sparse in some basis of $\mathbb{R}^n$ with only $O(k\log(n/k))$ linear projections and using low-complexity recovery algorithms [1, 2]. As such, compressive sensing has been proposed for myriad applications across signal processing. In addition to the *recovery* of low-dimensional signals, it is natural to consider other information-processing tasks in the compressive domain. Here we consider *classification* in the compressive domain. Rather than a signal-estimation problem, ours is a hypothesis-testing problem: How well can a classifier discern the signal class from a limited number of compressive measurements?

In this paper we study compressive *subspace* classification. The signal of interest belongs to a k-dimensional subspace of $\mathbb{R}^n$, and the classifier intends to identify the subspace from m linear projections. This problem is related to, but distinct from, sparse support recovery. The support recovery task is to identify the k-dimensional *canonical* subspace (or some rotation thereof) to which the signal of interest belongs; the number of such subspaces is dictated by combinatorics. By contrast, in our problem the subspaces are arbitrary; thus a classification task may involve any number of classes.

We derive fundamental limits on compressive subspace classification. Given k, m, n, and the received signal-to-noise ratio (SNR), we characterize the number of classes that can be discriminated with high probability via information-theoretic machinery. As the several signal dimensions approach infinity, we bound the logarithm of the number of discernible classes via the mutual information between the class subspace and the compressive measurements. Somewhat surprisingly, a duality between subspace classification and communications over non-coherent multiple-antenna channels [3] allows us to exploit well-known results from the communications literature in proving our claim.

Owing to space considerations, we merely state our theoretical results and devote the rest of the paper to their empirical validation. The results as provide rules of thumb for the design of practical classifiers, prescribing the number of measurements required to distinguish a desired number of classes. To test this interpretation, we study face recognition, which can be cast as a subspace identification problem [4, 5], in the compressive domain. We find that our results provide rather accurate predictions regarding the number of measurements required to discriminate faces with high probability.

2. PROBLEM STATEMENT

2.1. Signal Model

We suppose a linear, noisy signal model. The classifier obtains the signal of interest, passed through a sensing matrix and corrupted by additive Gaussian noise:

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{z}, \tag{1}$$

where $\mathbf{\Phi} \in \mathbb{R}^{m\times n}$ is a fixed sensing matrix, $\mathbf{x} \in \mathbb{R}^n$ is the signal of interest, and $\mathbf{z} \in \mathbb{R}^m$ is white Gaussian noise with covariance matrix $1/\mathrm{SNR}\cdot\mathbf{I}_{m\times m}$. We suppose $m \leq n$ throughout. We call n the *ambient signal dimension* and m the *number of measurements*. To ensure that the SNR remains meaningful, we impose energy constraints on both the signal of interest and the sensing matrix:

$$\|\mathbf{\Phi}\|_2^2 \leq 1, E[\|\mathbf{x}\|^2] \leq m,$$

where $\|\cdot\|_2$ is the ℓ_2 operator norm.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

We further suppose that $\mathbf{x}$ belongs to a k-dimensional subspace of $\mathbb{R}^n$. We define a *classification problem* as a collection of L subspaces, which we denote by an L-tuple of orthonormal matrices

$$\mathcal{P} = \{\mathbf{U}_1, \cdots, \mathbf{U}_L\}, \tag{2}$$

where each $\mathbf{U}_l \in \mathbb{R}^{n\times k}$ represents one of the subspaces to be discriminated. We suppose that the classifier knows $\mathcal{P}$ perfectly. Its task is to choose the correct $\mathbf{U}_l$ from the compressive measurements $\mathbf{y}$. In order to derive information-theoretic bounds, we further suppose that $\mathbf{x}$ follows a zero-mean Gaussian distribution with covariance $\mathbf{U}_l\mathbf{U}_l^T$. Equivalently, the signal of interest can be written as

$$\mathbf{x} = \mathbf{U}_l\mathbf{h}, \tag{3}$$

where $\mathbf{h} \in \mathbb{R}^k$ is distributed according to $\mathcal{N}(\mathbf{0}, \mathbf{I}_{k\times k})$. Then, the received signal can be written

$$\mathbf{y} = \Phi\mathbf{U}_l\mathbf{h} + \mathbf{z}. \tag{4}$$

2.2. Classification Capacity

A classifier is a mapping $\mathcal{C} : \mathbb{R}^m \to \{1, \ldots, L\}$ that takes as input the measurement $\mathbf{y}$ and returns as output an estimate $\hat{l}$ of the class associated with $\mathbf{x}$. Let the classes have uniform prior probability, and let the probability of misclassification be denoted $P_e = \Pr(\hat{l} \neq l)$. When the dimensions n, m, and k, and the number of classes L go to infinity, we can characterize via Shannon theory the regimes in which $P_e \to 0$. In particular, we prove limits on how fast the number of classes L can grow with probability of error going to zero.

Consider a sequence of classification problems $\mathcal{P}_m$, indexed by the number of measurements m. Each P_m may have a different value for n, m, k, and L. By analogy with communications problems, we can define the "rate" of the sequence, or the log-cardinality normalized by m:

$$\rho = \lim_{m\to\infty} \frac{\log(L)}{m}.$$

Furthermore, let n and k scale linearly with m, or

$$\lim_{m\to\infty} \frac{n}{m} = \nu, \ \lim_{m\to\infty} \frac{k}{m} = \kappa,$$

for $\nu \geq 1$ and $\kappa < 1$. With this parameterization, we can describe the error performance in the limit of large dimension, expressed in terms of ρ, ν, κ, and the SNR. We say that a normalized log-cardinality ρ is *achievable* if there exists a sequence $\mathcal{P}_m$ such that $P_e \to 0$ as $m \to \infty$. We call the supremum over achievable log-cardinalities ρ the *classification capacity*, and we denote it by $C(\nu, \kappa, \mathrm{SNR})$.

By Fano's inequality, we can upper bound the classification capacity by $\lim_{m\to\infty} I(\mathbf{U}; \mathbf{y})/m$, the normalized mutual information between the subspaces and the measurement $\mathbf{y}$. As long as ρ is greater than this quantity, the probability of error is bounded away from zero. While on its face it appears difficult to compute this mutual information, we identify a duality with non-coherent multiple-antenna communications channels that permits a relatively straightforward computation.

3. CAPACITY BOUNDS

Here we present bounds on the classification capacity. We omit the proofs, instead providing the intuition behind our analysis, which is rooted in a duality between the classification of Gaussian signals and communications over MIMO channels, which we describe in a previous work [6]. To illustrate this duality, we first briefly review the non-coherent MIMO channel as studied in [3]. It consists of a transmitter having N antennas, a receiver having T antennas, and an unknown, i.i.d. complex Gaussian channel matrix $\mathbf{H} \in \mathbb{C}^{T\times N}$ unknown to transmitter and receiver and persisting for M symbol times. The signal model is

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{Z}, \tag{5}$$

where $\mathbf{Y}$ and $\mathbf{Z}$ are $T \times M$ matrices, and $\mathbf{X}$ is $N \times M$. The noise signal $\mathbf{Z}$ is, as before, i.i.d. AWGN. The transmit signal $\mathbf{X}$ is subject to the usual energy constraint.

Equation (5) is similar to (4). In fact, taking the transpose of (4), we see that the signal model for classification is nearly identical to that of communications over the non-coherent MIMO channel:

$$\mathbf{y}^T = \mathbf{h}^T\mathbf{U}_l^T\Phi^T + \mathbf{z}^T. \tag{6}$$

The preceding equation nearly models a non-coherent MIMO channel having k transmit antennas, one receive antenna, and a coherence time of m. The matrix $\mathbf{U}_l$, which the classifier intends to recover, takes the place of the codeword. Just as the codeword in the non-coherent MIMO channel passes through an unknown channel matrix $\mathbf{H}$, the matrix $\mathbf{U}_l$ is transformed by the unknown driving process $\mathbf{h}$. Thus there is an equivalence between identifying the class subspace driven by an unknown Gaussian process and recovering a codeword transformed by an unknown Gaussian channel matrix. Important differences remain—for example, the classification signal model is real-valued, whereas the MIMO channel is complex, and the measurement matrix Φ has no direct analogue in the MIMO channel—but we can leverage much of the theoretical machinery to prove bounds on the classification capacity.

In [3] it is observed that the near-optimal distribution of codewords is an isotropic packing of subspaces across the Grassmann manifold. That is, the distribution that approximately maximizes the mutual information is uniform over the manifold of low-dimensional subspaces. The receiver discerns codewords by detecting the subspace to which the received signal belongs. Similarly, the optimum distribution of

the matrices $\mathbf{U}_l$ is uniform over the Grassmann manifold of k-dimensional subspaces of $\mathbb{R}^n$. Using this optimal distribution, one can show that the mutual information between the subspaces and the received signal is

$$I(\mathbf{U};\mathbf{y}) = \frac{m-k}{2}\log(\mathrm{SNR}) + O(m). \tag{7}$$

From this fact, we can state the main result of this paper

Theorem 1 *The classification capacity is*

$$C(\nu,\kappa,\mathrm{SNR}) = \frac{1-\kappa}{2}\log(\mathrm{SNR}) + O(1). \tag{8}$$

The converse part is a consequence of Fano's inequality and the above computation of $I(\mathbf{U};\mathbf{y})$. Achievability is proven by direct analysis of the misclassification probability over an ensemble of randomly-constructed subspaces. The proof will be presented in full in a later work.

The upshot of Theorem 1 is that if the number of classes grows faster than $\mathrm{SNR}^{(m-k)/2}$, the misclassification error is bounded away from zero, and if the number of classes grows slower, there exists a sequence of classification problems for which the misclassification error goes to zero. Naturally, for finite dimension there is not a hard threshold dividing success from failure. Nevertheless, these results provide design intuition whose effectiveness we study in the next section. Much as a communications engineer should take care to signal at rates below capacity, a system designer should take sufficiently many measurements that the classification capacity is not exceeded. Fortunately, the capacity result suggests that the number of discernible classes grows exponentially in m. For a fixed target cardinality, one usually does not need to take too many more measurements in order to increase sufficiently the classification capacity.

4. FACE RECOGNITION

Here we explore the correspondence between the theoretical results derived in the previous section and face recognition. While faces have been studied extensively, an interesting—and fruitful for our purposes—line of research transforms the face recognition task into a subspace classification one. It considers images of faces in which the orientation relative to the camera remains fixed, but the illumination varies. Supposing the faces themselves to be approximately convex and to reflect light according to Lambert's law, it was shown via spherical harmonics that the set of images of an individual face is well approximated by a nine-dimensional subspace [4]. Thus, discriminating between the 9D subspaces is sufficient to classify faces.

Given this linear characterization, we examine how well one can classify using compressive measurements, and whether the theory described in the previous section can inform system design. For our study, we use 38 cropped faces from the Extended Yale Face Database B, described in [5, 7]. For each face, the database contains a few dozen greyscale photographs, each having 32,256 pixels, taken under a variety of illumination conditions as shown in Figure 1. We vectorize these images, pass them through a measurement matrix Φ, and classify them as described below.

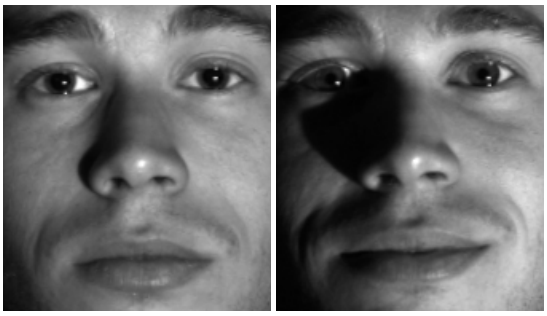

Fig. 1. Two sample images from the Extended Yale Face Database B. These images are of the same face, but are taken under different illumination conditions.

4.1. Compressive 9PL

We classify the faces according to a compressive variation on the *nine points of light* (9PL) algorithm presented in [5]. The 9PL algorithm is a low-complexity, low-training classifier inspired by the theoretical results of [4]. It involves both training and subspace classification. For training, nine images of each face are chosen to form a basis for the associated 9D subspace. These images are chosen to have near-optimum illumination patterns; the details are given in [5]. Consistent with the previous sections, let $\mathbf{U}_l \in \mathbb{R}^{32,256\times 9}$ be an orthonormal representation of the basis for face l.

Then, we classify an image according to the minimum-norm residual. We choose the rows of Φ as the M principal components of the training data. Then, for image $\mathbf{x}$, the classifier operates on the signal $\mathbf{y} = \Phi\mathbf{x}$, comparing the projection onto each subspace and choosing the class with the smallest projection error:

$$\hat{l} = \arg\min_l \left\|\mathbf{y} - \Pi_{\Phi\mathbf{U}_l}\mathbf{y}\right\|, \tag{9}$$

where $\Pi_{\mathbf{A}}$ is the orthogonal projection onto the column space of $\mathbf{A}$.

In Figure 2 we plot the misclassification probability as a function of m and for L ranging from 2 to 38. We randomly choose L faces, compute the subspaces via 9PL, and average the probability of misclassification error over 1000 random samplings from the classes. While we do not label each curve, it is easy to see that the misclassification probability increases with L and decreases with m. However, even for large m the error probability remains bounded away from

zero. Noise, non-Lambertian reflectances, shadows due to the non-convexity of real faces, misestimation of subspaces, etc., all result in a model mismatch, which produces an error floor.

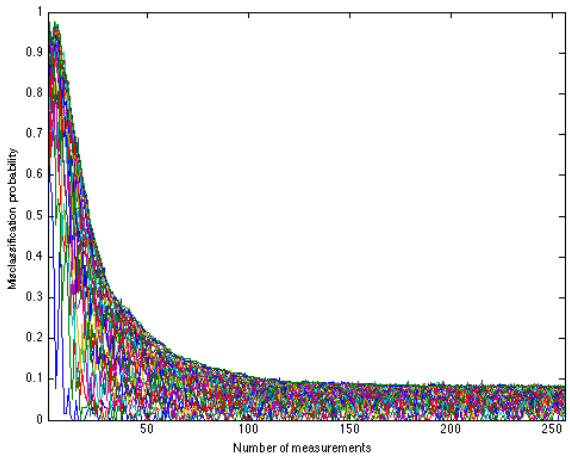

Fig. 2. Misclassification probability as a function of m, for L ranging from 1 to 38.

4.2. Agreement with Theory

Finally, we examine how well the rules of thumb derived in the previous sections predict the performance seen here. In order to do so, we need to translate the ostensibly noise-free images into the noisy signal model of (1). Because there is no additive noise explicit in the images, we compute the SNR empirically according to the average projection error. We project each compressed image onto the span of $\Phi\mathbf{U}_l$ and take the projected squared norm as the signal power and the squared residual norm as the noise power. Somewhat interestingly, the SNR *decreases* in m. As we take more measurements, we permit more non-idealities, and therefore more noise, into the received signal. However, this phenomenon impacts the results only slightly, since the number of classes grows exponentially in m, but only polynomially in the SNR.

Using the computed SNRs, we estimate the number of classes that Theorem 1 predicts can be discriminated reliably. We simply compute $\max\{1, \min\{\mathrm{SNR}^{(m-9)/2}, 38\}\}$. Naturally, this number grows quickly in m, and beyond $m = 11$ or $m = 12$, theory suggests that we ought to be able to discriminate all 38 of the faces with low probability of error. In Figure 3 we compare this prediction against the empirical performance of our classifier. Using the results shown in Figure 2, we compute, for each m, the maximum L for which the probability of error is less than 0.2.

The empirical performance is similar to theoretical prediction. As m increases past 9, the number of classes rises swiftly as predicted. After $m = 50$ or so, all 38 of the classes

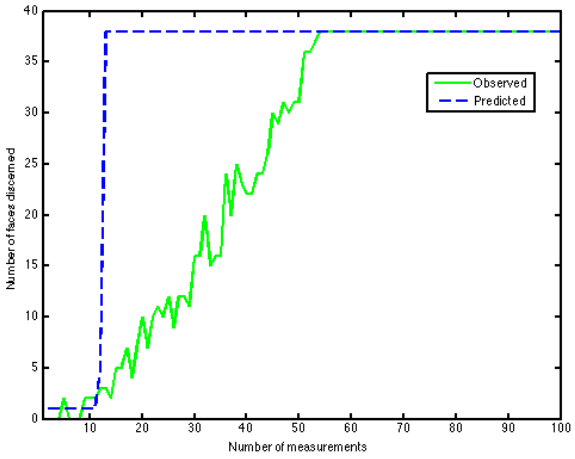

Fig. 3. Number of discernible classes as a function of m.

can be discriminated, and it is not advantageous to take more measurements. We do observe, however, that the transition is not as sharp as Theorem 1 predicts. Whereas the theoretical transition occurs over only 2-3 measurements, in practice the transition stretches out over 40 measurements. One could postulate a few explanations for this discrepancy. For example, the noise and the driving process are not Gaussian; the classifier is suboptimal and therefore may not achieve the classification capacity; and we are neglecting the $O(1)$ term in computing the number of discernible classes.

Instead, however, we conjecture that the discrepancy is due primarily to the energy distribution over the 9D subspaces. Like most natural images, the spectrum of the face images follows a power-law distribution. Thus each additional measurement captures a smaller marginal signal energy. Our modeling assumptions, however, suppose that the subspace is driven by a white process, so each additional dimension ought to capture an equal marginal signal energy. Therefore the improvement with m is somewhat smaller than predicted. A fruitful area for future research is a theoretical characterization of subspace classification when the modes are not excited equally.

5. CONCLUSION

We have presented information-theoretic limits on the performance of compressive classifiers over low-dimensional subspaces. These limits provide design intuition for practical compressive classification systems. This intuition is validated via face recognition algorithms, where we observe reasonably close correspondence between theory and practice. Future work includes the analysis when subspaces are excited according to a power-law distribution.

6. REFERENCES

[1] D. L. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[2] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[3] L. Zheng and D. N. C. Tse, "Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. Info. Theory*, vol. 48, no. 2, pp. 359–383, 2002.

[4] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 218–233, 2003.

[5] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[6] M. Nokleby, M. Rodrigues, and R. Calderbank, "Information-theoretic limits on the classification of Gaussian mixtures: Classification on the Grassmann manifold," in *Proc. Information Theory Workshop (ITW)*, 2013.

[7] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.