HOW TO FOCUS THE DISCRIMINATIVE POWER OF A DICTIONARY

William R. Carson[*]*, Miguel R. D. Rodrigues*[*], *Minhua Chen*[†]*, Lawrence Carin*[†] *and Robert Calderbank*[†]

[*] Instituto de Telecomunicações, Departamento de Ciência de Computadores, Universidade do Porto.
[†] Department of Electrical Engineering, Duke University.

ABSTRACT

This paper is motivated by the challenge of high fidelity processing of images using a relatively small set of projection measurements. This is a problem of great interest in many sensing applications, for example where high photodetector counts are precluded by a combination of available power, form factor and expense. The emerging methods of dictionary learning and compressive sensing offer great potential for addressing this challenge. Combining these methods requires that the signals of interest be representable as a sparse combination of elements of some dictionary. This paper develops a method that aligns the discriminative power of such a dictionary with the physical limitations of the imaging system. Alignment is accomplished by designing a projection matrix that exposes and then aligns the modes of the noise with those of the dictionary. The design algorithm is obtained by modifying an algorithm for designing the pre-filter to maximize the rate and reliability of a Multiple Input Multiple Output (MIMO) communications channel. The difference is that in the communications problem a source is being matched to a channel, whereas in the imaging problem a channel, or equivalently the noise covariance, is being matched to a source.

Our results shown that using the proposed communications design framework we can reduce reconstruction error between 20%, after only 20 projections of a 28×28 image, and 10% after 100 projections. Furthermore, we noticeably see the superior quality of the reconstructed images.

Index Terms— Low Resolution Imaging, Compressed Sensing, MIMO Communication, Precoder Design, Mode Alignment, Mutual Information

1. INTRODUCTION

Compressive sensing (CS) has received significant attention in recent years due to the impressive results that are promised for the reconstruction of a high-dimensional signal with a relatively small set of random projection measurements. The results hinge on the signal in question being sparse, that is the signal belongs to a low dimensional manifold although it lives in a much higher dimensional space. The optimization problem of designing the measurement matrix that minimizes the reconstruction error subject to unit norm columns has parallels with the communications theory problem of optimal precoder design subject to a power constraint.

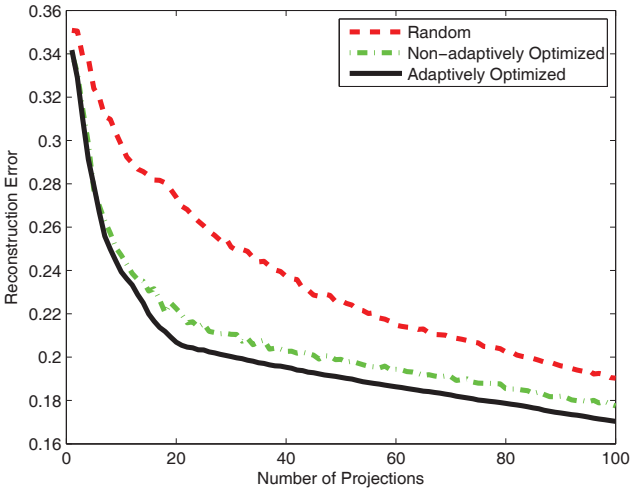
In a communications system, an objective that is well-studied is how to maximize the mutual information between the input and output of a channel; one instance is the design of the optimal precoder that adapts the signal to the channel, subject to a power constraint. The optimal solution for this problem has been characterized for Gaussian inputs [1], non-Gaussian inputs with a diagonal precoder [2] and non-Gaussian inputs with a general precoder [3]. Many of these results build on a fundamental result for one-dimensional signals [4], and its extensions to multivariate sources [5]. This result relates the derivative of the mutual information and the minimum mean squared error (MMSE), thus connecting information theory to estimation.

We will demonstrate how communication theory can be applied to great effect in an image processing context. Consider the reconstruction of the MNIST digit data in Fig. 1(a); training data is used to learn a dictionary that is represented by a Gaussian mixture model (GMM) which describes the data by a mixture of Gaussians living in (possibly) overlapping sub-spaces. Such models have been shown to accurately characterize true images and are well-accepted in the literature [6]. We then compare the performance in three separate scenarios. In the first scenario, we sense and reconstruct the data using a projection matrix with elements drawn iid from $\mathcal{N}(0,1)$, which has been shown to lead to good performance for sparse signals [7]. In the second scenario in Fig. 1(a), we design a one-shot projection matrix based on the information-theoretic results in this paper which allow us to represent explicitly, in terms of fixed point equations, the optimal projection matrix for a general multivariate. In the third scenario, we are able to adaptively modify the measurement matrix using information from previous observations, again using the proposed framework. We see from the qualitative image results in Fig. 1(a) and the quantitative MSE results in Fig. 1(b) that image reconstruction based on the information theoretic principles is superior to reconstruction based on random projection. Note

[*]The work of W. R. Carson and M. R. D. Rodrigues was supported by the Fundação para a Ciência e a Tecnologia through research projects PTDC/EEA-TEL/100854/2008 and CMU-PT/SIA/0026/2009.

[†]The work of M. Chen, L. Carin and R. Calderbank was supported in part by NSF under Grant DMS-0914892, by ONR under Grant N00014-08-1-1110 and by DARPA under the KeCom Program.

(a) The first column is the ground truth, subsequent columns show reconstructions for 5, 10, 20, 50 and 100 projections of a random projection matrix, an optimized one-shot (non-adaptive) projection matrix and an optimized adaptive projection matrix, respectively.



(b) Reconstruction error.

Fig. 1. Reconstruction of MNIST digit data.

that all CS recovery results in Figs. 1(a) and 1(b) employ a learned GMM. If one were to instead simply use conventional CS with sparsity in an orthonormal basis (wavelets), the signal recovery is much worse [6].

In the rest of this paper we outline the theorems that justify the design principles behind these gains. The principle of optimizing the performance of imaging systems via mutual information is not a new topic, the novelty in this paper is that the design principles presented here are valid for *any multivariate source distribution* (in Fig. 1(a) the images are modelled by a Gaussian mixture model (GMM) with low-rank covariances [6]).

2. SYSTEM MODEL

We consider a general discrete-time noisy projection $\mathbf{y} \in \mathbb{R}^k$, which is given by:

$$\mathbf{y} = \sqrt{\mathsf{g}} \cdot \mathbf{M}\,\mathbf{D}\,\mathbf{x} + \mathbf{w} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the source vector, the matrices $\mathbf{M} \in \mathbb{R}^{k\times m}$ and $\mathbf{D} \in \mathbb{R}^{m\times n}$ represent the measurement/projection matrix and the dictionary, respectively, and $\mathbf{w} \in \mathbb{C}^k$ represents a zero-mean Gaussian noise with covariance $\mathbf{\Sigma_w}$.

Without loss of generality, we assume that $\mathbf{x}$ has zero-mean and covariance $\mathbf{\Sigma_x} = \mathbf{I}$. We shall design the measurement matrix throughout via its singular value decomposition (SVD) $\mathbf{M} = \mathbf{U_M}\mathbf{\Lambda_M}\mathbf{V_M^\intercal}$ where $\mathbf{\Lambda_M} = \mathsf{diag}\left(\{\sqrt{\lambda_{M_i}}\}\right)$, and the eigenvalue decomposition (EVD) of the key model matrices: the positive definite noise covariance $\mathbf{\Sigma_w} = \mathbf{U_w}\mathbf{\Lambda_w}\mathbf{U_w^\intercal}$, where $\mathbf{\Lambda_w} = \mathsf{diag}\left(\{\lambda_{w_i}\}\right)$, and the positive semi-definite matrix $\mathbf{D}\mathbf{\Sigma_x}\mathbf{D}^\intercal = \mathbf{U_D}\mathbf{\Lambda_D}\mathbf{U_D^\intercal}$ where $\mathbf{\Lambda_D} = \mathsf{diag}\left(\{\lambda_{D_i}\}\right)$ and the positive semi-definite matrix $\mathbf{D}\mathbf{E}\mathbf{D}^\intercal = \mathbf{U_E}\mathbf{\Lambda_E}\mathbf{U_E}^\intercal$, where $\mathbf{\Lambda_E} = \mathsf{diag}\left(\{\lambda_{E_i}\}\right)$ and $\mathbf{E}$ is the MMSE matrix associated with this model:

$$\mathbf{E} = \mathbb{E}\left\{(\mathbf{x} - \mathbb{E}\left\{\mathbf{x}|\mathbf{y}\right\})(\mathbf{x} - \mathbb{E}\left\{\mathbf{x}|\mathbf{y}\right\})^\intercal\right\} \tag{2}$$

It is important to note that the MMSE matrix is also a function of the measurement matrix. We consider the design of the measurement matrix that maximizes the mutual information between the source vector and the measurement vector, i.e., we pose the optimization problem:

$$\max_{\mathbf{M}}\ \mathsf{I}\left(\mathbf{x}; \sqrt{\mathsf{g}}\cdot\ \mathbf{M}\,\mathbf{D}\,\mathbf{x} + \mathbf{w}\right)\ \text{s.t. tr}\left(\mathbf{M}\mathbf{M}^\intercal\right) \le \mathsf{P}. \tag{3}$$

3. OPTIMAL PROJECTION MATRIX: MULTIVARIATE GAUSSIAN SOURCE

When we consider the design of the measurement matrix for a multivariate circularly symmetric real Gaussian source, we are able to take advantage of the well-known expressions for mutual information and the MMSE matrix:

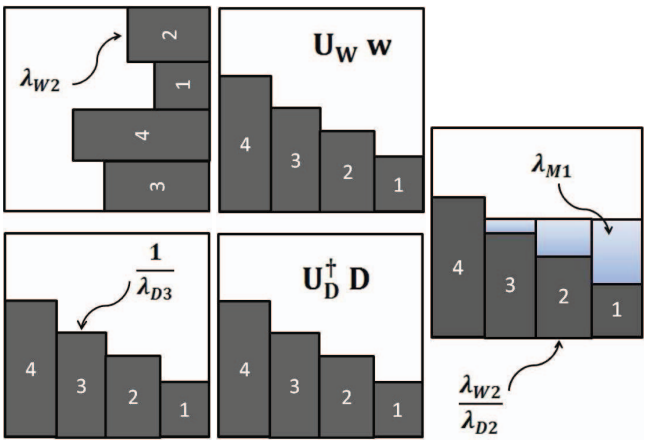
$$\mathsf{I}\left(\mathbf{x};\mathbf{y}\right) = \frac{1}{2}\log\det\left(\mathbf{I} + \mathsf{g}\cdot\mathbf{\Sigma_w}^{-\frac{1}{2}}\mathbf{M}\mathbf{D}\mathbf{\Sigma_x}\mathbf{D}^\intercal\mathbf{M}^\intercal\mathbf{\Sigma_w}^{-\frac{1}{2}}\right) \tag{4}$$

$$\mathbf{E} = \left(\mathbf{\Sigma_x}^{-1} + \mathsf{g}\cdot\mathbf{D}^\intercal\mathbf{M}^\intercal\mathbf{\Sigma_w}^{-1}\mathbf{M}\mathbf{D}\right)^{-1}, \tag{5}$$

where $\mathbf{I}$ is the identity matrix. The optimization problem in (3) is solved by the following theorem,which capitalizes on the relationship between the gradient of the mutual information and the MMSE matrix [5], and also builds upon the contributions in [8] and [9].

Theorem 1. *The optimal measurement matrix that solves* (3) *for a multivariate circularly symmetric real Gaussian source with mean* $\mathbb{E}\left\{\mathbf{x}\right\} = \mathbf{0}$ *and covariance* $\mathbb{E}\left\{\mathbf{x}\mathbf{x}^\intercal\right\} = \mathbf{I}$ *is:*

$$\mathbf{M}^\star = \mathbf{U_w}\mathsf{diag}\left(\sqrt{\left(\frac{1}{\eta} - \frac{1}{\mathsf{g}}\cdot\frac{\lambda_{w_i}}{\lambda_{D_i}}\right)^+}\right)\mathbf{U_D^\intercal} \tag{6}$$



(a) Exposing modes (b) Aligning and ordering modes (c) Waterfilling power allocation

Fig. 2. Diagrammatic representation of the actions of the optimal projection matrix.

where η is such that $\mathsf{tr}(\mathbf{M}\mathbf{M}^\mathsf{T}) = \mathsf{P}$ and the eigenvalues of $\mathbf{D}\mathbf{D}^\mathsf{T}$ and $\mathbf{\Sigma_w^{-1}}$ are arranged in descending order, i.e., $\lambda_{D1} \geq \lambda_{D2} \geq \cdots > 0$ and $\lambda_{w1}^{-1} \geq \lambda_{w2}^{-1} \geq \cdots > 0$.

It is also important to reflect on the nature of the operations carried out by the optimal projection matrix embodied in Theorem 1 which we present diagrammatically. In Fig. 2(a) we depict that the optimal projection procedure is such that it exposes the modes of the noise and the modes of the dictionary. Then, in Fig. 2(b) we illustrate that the procedure performs an alignment and ordering operation whose purpose it to optimally match the modes of the noise to the modes of the dictionary. Finally, Fig. 2(c) shows the, aptly described, waterfilling power allocation policy [1].

Proof. The solution follows from the Karush-Kuhn-Tucker (KKT) optimality conditions [10]:

$$\eta \cdot \mathbf{M}^\star = \mathsf{g}\, \mathbf{\Sigma_w^{-1}} \mathbf{M}^\star \mathbf{D}\, \mathbf{E}^\star \mathbf{D}^\mathsf{T} \tag{7}$$

where $\eta \geq 0$ and a superscript star $(\star)$ denotes the optimal solution, for example, $\mathbf{E}^\star$ is the MMSE matrix associated with the optimal measurement matrix $\mathbf{M}^\star$, i.e., $\mathbf{E}^\star = \mathbf{E}\left(\mathbf{M}^\star\right)$.

Post-multiplying the fixed-point equation in (7) by $\mathbf{M}^{\star\mathsf{T}}$, we see that the resultant matrix is symmetric and is composed of two positive semi-definite matrices that commute; this implies that they are simultaneously diagonalizable by the unitary matrix $\mathbf{U}_\mathbf{M}^*$. Subsequently, one can also infer that the unitary matrix $\mathbf{V}_\mathbf{M}^*$ diagonalizes not only the positive semi-definite matrix $\mathbf{D}\, \mathbf{E}^\star \mathbf{D}^\mathsf{T}$, but also $\mathbf{D}\mathbf{\Sigma_x}\mathbf{D}^\mathsf{T}$, i.e.,

$$\mathbf{U}_\mathbf{M}^\star = \mathbf{U_w}\mathbf{\Pi}_\mathbf{U}^\star \tag{8}$$

$$\mathbf{V}_\mathbf{M}^\star = \mathbf{U_D}\mathbf{\Pi}_\mathbf{V}^\star \tag{9}$$

where $\mathbf{\Pi}_\mathbf{U}^\star$ and $\mathbf{\Pi}_\mathbf{V}^\star$ are permutation matrices[1] (c.f., [9]).

Using these optimal unitary matrices, the matrix optimization problem reduces to a set of scalar optimization problems that are concave in $\lambda_{M_i}, i = 1, \ldots, n$ for a fixed permutation matrix; directly from the KKT conditions the unique solution to the optimization problem is then:

$$\lambda_{M_i}^\star = \begin{cases} 0, & \eta \geq \mathsf{g} \cdot \frac{\lambda_{D_i}}{\lambda_{w_{\pi^\star(i)}}} \\ \frac{1}{\eta} - \frac{1}{\mathsf{g}} \cdot \frac{\lambda_{w_{\pi^\star(i)}}}{\lambda_{D_i}}, & \eta < \mathsf{g} \cdot \frac{\lambda_{D_i}}{\lambda_{w_{\pi^\star(i)}}} \end{cases} \tag{10}$$

where η is such that $\sum \lambda_{M_i}^\star = \mathsf{P}$ and $\{\pi^\star(i)\}$ denotes the permutation induced by the matrix $\mathbf{\Pi}^\star = \mathbf{\Pi}_\mathbf{V}^\star \mathbf{\Pi}_\mathbf{U}^{\star\mathsf{T}}$.

It is important to emphasize that one of the key ingredients in the measurement process relates to the alignment of the source modes to the filter modes. Without loss of generality, we arrange the noise eigenmodes in order of *ascending* eigenvalue. For two source and two noise modes, denoted 2×2, it is fairly straight-forward to prove that arranging the source eigenmodes in *descending* order is optimal.

Consider an $N \times N$ scenario where the source eigenmodes are sorted in descending order except for two elements, indexed by i and j. These two elements are allocated power $\mathsf{P}_2 = \mathsf{P}_i + \mathsf{P}_j$ of the total power P. We view these two unordered elements as a 2×2 scenario with power P_2, exchanging these two elements and allocating P_2 optimally between the two of them increases the mutual information[2] If we start from the case when the source eigenmodes are in descending order, we can obtain any permutation by exchanging adjacent elements to be in ascending order, this will always decrease the maximum mutual information and therefore $\mathbf{\Pi}^\star = \mathbf{I}$.

□

4. OPTIMAL PROJECTION MATRIX: GENERAL MULTIVARIATE SOURCE

We now consider the design of an optimal measurement matrix for a general multivariate source; this could be a model where all signals live on the same sub-space, as is often used in communications, or it can represent a union of sub-spaces model where signals live on different sub-spaces, such as the GMM. The optimal solution for this scenario is based on a recent result in information theory relating the gradient of mutual information to the minimum mean squared error [5]. It is stated in the following theorem:

Theorem 2. *The optimal measurement matrix that solves* (3) *for a general multivariate source with mean $\mathbb{E}\{\mathbf{x}\} = \mathbf{0}$ and covariance $\mathbb{E}\{\mathbf{x}\mathbf{x}^\mathsf{T}\} = \mathbf{\Sigma_x}$ is (without loss of generality):*

$$\mathbf{M}^\star = \mathbf{U_w}\mathbf{\Pi}^\star \mathsf{diag}\left(\sqrt{\lambda_{M_i}^\star}\right) \mathbf{U}_\mathbf{E}^{\star\mathsf{T}} \tag{11}$$

[1] The general solution for both matrices includes a post-multiplication by diagonal matrices with unit modulus diagonal elements, however, we observe that the MMSE matrix and mutual information are insensitive to them.

[2] This is a lower bound on the gain since the other $N-2$ elements are no longer necessarily allocated optimal power.

where

$$\lambda_{M_i} = \begin{cases} 0, \\ \quad \eta \geq \mathsf{g}\lambda_{W_{\pi^\star(i)}} \mathsf{mmse_i}\left(\ldots, \lambda^\star_{M_{i-1}}, 0, \lambda^\star_{M_{i+1}}, \ldots\right) \\ \mathsf{mmse_i}^{-1}\left(\ldots, \lambda^\star_{M_{i-1}}, \frac{\eta}{\mathsf{g}\cdot\lambda_{W_{\pi^\star(i)}}}, \lambda^\star_{M_{i+1}}, \ldots\right), \\ \quad \eta < \mathsf{g}\lambda_{W_{\pi^\star(i)}} \mathsf{mmse_i}\left(\ldots, \lambda^\star_{M_{i-1}}, 0, \lambda^\star_{M_{i+1}}, \ldots\right) \end{cases} \tag{12}$$

$\{\pi^\star(i)\}$ denote the permutation induced by the optimal permutation matrix $\mathbf{\Pi}^\star$, η is such that $\sum \lambda_{M_i} = \mathsf{P}$, the i-th element of the MMSE matrix is

$$\mathsf{mmse_i}\left(\{\lambda_{M_j}\}\right) := \mathbb{E}\left\{|\mathbf{s}'_i - \mathbb{E}\left\{\mathbf{s}'_i|\mathbf{y}'\right\}|^2\right\} \tag{13}$$

where $\mathbf{y}' = \sqrt{\mathsf{g}}\cdot\mathbf{\Pi}^\star\mathbf{\Lambda}_\mathbf{W}^{-\frac{1}{2}}\mathbf{\Pi}^{\star\intercal}\mathbf{\Lambda}_\mathbf{M}\,\mathbf{s}' + \mathbf{\Pi}^\star\mathbf{\Lambda}_\mathbf{w}^{-\frac{1}{2}}\mathbf{U}_\mathbf{w}^\intercal\mathbf{w}$, $\mathbf{s}' = \mathbf{U}^\star_\mathbf{E}\mathbf{D}\mathbf{x}$, and its inverse with respect to the composition of functions, in the argument $\lambda^\star_{M_i}$ with $\lambda_{M_j}, \forall j \neq i$, fixed is $\mathsf{mmse_i}^{-1}(\ldots, \lambda_{M_{i-1}}, (\cdot), \lambda_{M_{i+1}}, \ldots)$.

Proof. This details of the proof are omitted due to space considerations, however, this is a generalization of Theorem 2 to include a general source covariance and builds on several contributions [8] [3]. □

5. OPTIMAL ADAPTIVE PROJECTION MATRIX DESIGN

The designs described thus far deal with one-shot designs for the measurement matrix. In several scenarios, for example, in the repetitive slices taken by MRI machines, it may be feasible to design the rows of the projection matrix sequentially, taking into account measurements from the previous projections. At each iteration, the problem can be seen as a one-shot problem with a new priori distribution on the source and therefore we apply results in Section 3 and Section 4 at each stage. There are, however, certain simplifications due to the reduced dimensionality of the design.

Firstly, the noise is now a scalar and each row has unit-norm, therefore $\mathbf{U}^\star_\mathbf{M} = 1$ and $\lambda^\star_\mathbf{M} = 1$. The problem reduces to selecting a vector from the eigenvectors of the MMSE matrix, which are the only vectors that satisfy the KKT conditions. For the Gaussian source, the optimal row vector $\mathbf{m}^\star$ is the eigenvector associated with the largest eigenvalue of the source covariance.

6. CONCLUSIONS

We observe that the design principle of maximizing mutual information leads to deterministic projection matrices for which MMSE performance is superior to that of random Gaussian adaptation matrices. In particular, we are able to provide design principles for the optimal projection matrix for a general multivariate source. We showed that the optimal measurement procedure exposes the modes of the noise and the modes of the (optimal) MMSE matrix, then performs an alignment operation whose purpose it to optimally match the modes of the noise to the modes of the MMSE matrix (or, in the multivariate Gaussian source scenario, the modes of the source covariance). Finally, it carries out a generalized mercury-waterfilling power allocation operation.

The visual depiction of the achievable gains were demonstrated on the MNIST digital data set, where the implementation of the design principles were able to achieve better reconstruction, in terms of both the perceived clarity of the images and the reconstruction error.

7. REFERENCES

[1] E. Telatar, "Capacity of multi-antenna Gaussian channels," in *AT&T-Bell Labs, Tech. Rep.*, 1995.

[2] A. Lozano, A. Tulino, and S. Verdú, "Optimum power allocation for parallel Gaussian channels with arbitrary input distributions," in *IEEE Trans. Inf. Theory*, July 2006, pp. 3033–3051.

[3] F. Pérez-Cruz, M. R. Rodrigues, and S. Verdú, "MIMO Gaussian channels with arbitrary inputs: Optimal precoding and power allocation," in *IEEE Trans. Inf. Theory*, Mar. 2010, pp. 1070–1084.

[4] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," in *IEEE Trans. Inf. Theory*, Apr. 2005, pp. 1261–1282.

[5] D. Palomar and S. Verdú, "Gradient of mutual information in linear vector Gaussian channels," in *IEEE Trans. Inf. Theory*, Jan. 2006, pp. 141–154.

[6] M. Chen, J. Silva, J. Paisley, C. Wang, D. Dunson, and L. Carin, "Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds," in *IEEE Trans. Signal Process.*, vol. 58, no. 12, Dec. 2010, pp. 6140 –6155.

[7] E. Candes and M. Wakin, "An introduction to compressive sampling," in *IEEE Signal Process. Mag.*, vol. 25, no. 2, 2008, pp. 21–30.

[8] M. Payaró and D. P. Palomar, "On optimal precoding in linear vector Gaussian channels with arbitrary input distribution," in *IEEE ISIT '09*, July 2009, pp. 1085–1089.

[9] M. Lamarca, "Linear precoding for mutual information maximization in MIMO systems," in *ISWCS '09*, Sept. 2009, pp. 26–30.

[10] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.