Communications-Inspired Projection Design with Application to Compressive Sensing

William R. Carson, Minhua Chen, Miguel R. D. Rodrigues,
Robert Calderbank and Lawrence Carin

***Abstract*—We consider the recovery of an underlying signal $\mathrm{x} \in \mathbb{C}^m$ based on projection measurements of the form $\mathrm{y} = \mathrm{M}\mathrm{x} + \mathrm{w}$, where $\mathrm{y} \in \mathbb{C}^\ell$ and w is measurement noise; we are interested in the case $\ell \ll m$. It is assumed that the signal model $p(\mathrm{x})$ is known, and $\mathrm{w} \sim \mathcal{CN}(\mathrm{w}; 0, \Sigma_{\mathrm{w}})$, for known Σ_{w}. The objective is to design a projection matrix $\mathrm{M} \in \mathbb{C}^{\ell \times m}$ to maximize key information-theoretic quantities with operational significance, including the mutual information between the signal and the projections $\mathcal{I}(\mathrm{x};\mathrm{y})$ or the Rényi entropy of the projections $\mathbf{h}_\alpha(\mathrm{y})$ (Shannon entropy is a special case). By capitalizing on explicit characterizations of the gradients of the information measures with respect to the projections matrix, where we also partially extend the well-known results of Palomar and Verdú from the mutual information to the Rényi entropy domain, we unveil the key operations carried out by the optimal projections designs: *mode exposure* and mode alignment. Experiments are considered for the case of compressive sensing (CS) applied to imagery. In this context, we provide a demonstration of the performance improvement possible through the application of the novel projection designs in relation to conventional ones, as well as justification for a fast online projections design method with which state-of-the-art adaptive CS signal recovery is achieved.**

I. Introduction

Compressive sensing (CS) [1], [2] has recently emerged as an important area of research in image sensing and processing. Compressive sensing has been particularly successful in multidimensional imaging applications, including magnetic resonance [3], spectral imaging [4], [5] and video [6], [7]. Conventional sensing systems typically first acquire data in an uncompressed form (*e.g.*, individual pixels in an image) and then perform compression subsequently, for storage or communication. In contrast, CS involves acquisition of the data in an already compressed form, reducing the quantity of data that need be measured in the first place. In CS the underlying signal to be measured is projected onto a set of vectors [8], [2], and one must perform an inverse problem to recover the underlying signal of interest.

There are two hallmarks of the original CS theory. First, the projection vectors were usually constituted uniformly at random. Second, the underlying signal model used to regularize the inverse problem was based on the assumption that the underlying signal could be sparsely represented in terms of an orthonormal basis or frame. However, even in some of the early CS studies, it was recognized that improved performance could be achieved with projection vectors designed to the underlying signal of interest [9], [10], [11], rather than using random projections. Further, it has recently been recognized that a signal model based upon sparsity is often overly primitive, and model-based CS [12], wherein improved signal models are employed, may yield improved CS performance (high-quality signal recovery with fewer projection measurements). Signal models that have been considered include the Gaussian mixture model (GMM) [13], union-of-subspace models [14], and manifold models [15].

In this paper our goal is to design CS projection matrices ("measurement kernels") matched to a general statistical signal model. Specifically, if the underlying signal to be measured is $\mathrm{x} \in \mathbb{C}^m$, it is assumed that we have access to a general signal model, represented statistically by density function $p(\mathrm{x})$. Our objective is to design the projection matrix to maximize the mutual information between the underlying signal and the observed compressive measurements or to maximize the Rényi of the compressive measurements.

The key to the approach considered in this paper is the realization that the projection-design problem for CS systems (subject to a power constraint) exhibits parallels with the precoder design problem for multiple-input–multiple-output (MIMO) communications systems: in the communications problem a source is being matched to a channel whereas in

W. R. Carson and M. R. D. Rodrigues are with the Instituto de Telecomunicações, Departamento de Ciência de Computadores, Universidade do Porto, Portugal. Their work was supported by the Fundação para a Ciência e a Tecnologia through the research project PTDC/EEA-TEL/100854/2008.

M. Chen, R. Calderbank and L. Carin are with Department of Electrical Engineering, Duke University, USA. Their work was supported in part by NSF under Grant DMS-0914892, by ONR under Grant N00014-08-1-1110, and by DARPA under grant N66001-11-1-4002 as part of the KeCom Program.

A part of the material presented in this work has been published in IEEE ICASSP 2012.

CS a channel, or equivalently the noise covariance, is being matched to the source. This link has also been recognized recently by Schnitter [16], who has provided projections designs for sources modelled by multivariate Gaussian distributions, as well as by Carson *et al.* [17], who have also provided designs for general multivariate source distributions. With the precoder design problem exhibiting a long tradition in the information theory and communications field, this link also provides the means to translate, with appropriate modifications, much of the design know-how and experience from the communications domain to CS.

The traditional problem of precoder design for MIMO Gaussian channels has been drawing on various performance metrics relevant for data communications. Common precoder design approaches aim to maximize the system signal-to-noise ratio (SNR) and the system signal-to-interference-plus-noise ratio (SINR) [18], [19] or minimize the system error probability [20], [21]. Another emerging precoder design approach imbued with operational significance is based on the maximization of the mutual information between the input and the output of the system [22], [23], [24], [25], [26]. This novel design principle has been shown to yield considerable rate gains in a variety of communications scenarios, due to the fact that, in addition to adapting to the channel characteristics, the designs also adapt to important features necessary to achieve high-rate reliable communications (the designs conform to the exact characteristics rather than only to the second-order statistics of the signaling scheme, as in traditional approaches (see [18], [19])). The basis of the emergence of the mutual information based designs have been fundamental connections between information theory and estimation theory, which have unveiled the interplay between mutual information and the minimum mean-squared error (MMSE) in scalar Gaussian channels [27] or mutual information and the MMSE matrix in vector Gaussian channels [28]. These results offer a means to bypass the absence of closed-form mutual information expressions for MIMO Gaussian channels driven by arbitrary (non-Gaussian) signaling schemes.

The operational significance of mutual information, which acts as the rationale for its use as the basis of a plethora of designs, is well known not only in data communications – it represents the highest reliable information transmission rate in a single-user channel driven by a specific signalling scheme – but also in other domains. For example, in classification problems mutual information relates (through bounds) to the Bayesian error probability of the classifier [29]; and, in regression problems mutual information relates (through bounds) to the reconstruction error [30].

We consider design of the measurement kernel based upon maximizing the mutual-information between the underlying signal $\mathbf{x}$ and the compressive measurement $\mathbf{y}$. We also consider design based upon maximizing the Rényi entropy of $\mathbf{y}$, where the latter represents a generalization with operational relevance [31]. The projection design will be implemented in practice using gradient descent, and we demonstrate that for a GMM signal model the gradient of Rényi entropy with respect to the design matrix may be expressed analytically, for a special parameter setting. Further, we recover the gradient of Shannon entropy as a special case of the Rényi result.

The article considers both theoretical results, which disclose key operations effected by the projection designs, as well as experimental results that demonstrate the merit of the approach as applied to a practical CS imaging problem. One key operation relates to the notion of *mode alignment* in mutual information based designs: the modes of the source, which depending on the source statistical model are given by the eigenvalues of the source covariance matrix or the eigenvalues of the source MMSE matrix, have to align with the modes (eigenvalues) of the noise covariance matrix as a means to improve performance. This role can also be conceptually appreciated by viewing the measurement kernel as a sieve that aligns relevant statistical features of the source to the statistical features of the noise, in order to disclose relevant information for reconstruction. The relevance of mode alignment, which is typically absent in communications problems[1], has also been recently unveiled in radar applications [32].[2] Overbridging the theoretical and practical results is also the formal justification of a low-complexity high-performance online projections strategy, the partial direction sensing method (PSD) [33], which brings together the main operational features of the optimal measurement kernel designs, including mode alignment.

The detailed contributions of the article include:

- Recognition that recent advances in communications, which relate to the design of precoders for MIMO communications channels, carry over to CS, leading to a communications-inspired kernel design framework for CS applications.
- Proposal of mutual information based offline – where a set of projections is optimized simul-

[1] This operation is absent in precoder designs for MIMO Gaussian channels driven by Gaussian inputs, due to the fact that the signal covariance is often taken to be white, but is present in precoder designs for MIMO Gaussian channels driven by non-Gaussian inputs. The role of a certain permutation operation in the precoder design is hinted at by Lamarca in [24].

[2] Note that Schniter [16] does not recognize the role of mode alignment due to the statistical assumptions about the source and noise covariances: this operation is not present when the source covariance is the identity matrix or when the noise covariance is also the identity matrix.

taneously – and online – where the individual projections are optimized sequentially – kernel designs. The article unveils key operations carried out by the optimal kernel designs for multivariate Gaussian sources and general multivariate sources, including the operations of source and noise modes exposure, mode weighting and mode alignment. Particular emphasis is given to the role of mode alignment as a means to improve further the reconstruction performance in compressive sensing applications.

- Proposal of Rényi entropy based kernel designs. The article also underlines some relations between the mutual information (or Shannon entropy) and the Rényi entropy based kernel constructions.
- Formal rationale for the PDS strategy [33], which is based on the operational insight unveiled by the theoretical characterizations of the optimal kernel designs.
- Partial generalizations of the I-MMSE identity from the mutual information (or Shannon entropy) to the Rényi entropy domain.
- A range of experimental results that illustrate the benefit of the novel measurement kernel designs in relation to the conventional random ones.

The remainder of the article is organized as follows. In Section II we briefly summarize the notation used throughout. Section III reviews the modeling and design approach, introducing key system assumptions. Section IV introduces the optimal kernel design based on the Shannon-based mutual information metric – this builds upon work on the communications field on precoder design for MIMO channels driven by Gaussian inputs and arbitrary inputs. Section V introduces the optimal kernel design based on the Rényi entropy metric, taking advantage of the closed-form expressions available for a GMM source. Section VI provides the body of evidence that demonstrates the performance improvement possible through the application of the projections designs put forth in previous sections. We consider examples based on offline kernel design, based upon the prior signal model, as well as online kernel design based upon sequential update of the posterior, all within the context of a GMM signal representation, which yields analytic CS inversion. Section VII draws the main conclusions. The Appendices contain proofs and supporting mathematical derivations.

II. Notation and Definitions

In the following text scalar quantities are denoted by italics, vectors are denoted by boldface lower case letters and matrices are denoted by boldface upper case letters. The projection of scalar x onto the nonnegative orthant is denoted $(x)^+ \triangleq \max(0, x)$. The superscript $(\cdot)^\star$ is used to denote an optimal solution and the superscripts $(\cdot)^T$, $(\cdot)^*$ and $(\cdot)^\dagger$ denote transpose, conjugate and conjugate transpose operators, respectively. The element in the i-th row and j-th column of the matrix $\mathbf{X}$ is denoted by $[\mathbf{X}]_{i,j}$. The trace of a matrix is denoted $\mathsf{tr}(\cdot)$. The diagonal matrix with diagonal elements given by either vector $\mathbf{x}$ or the diagonal elements of matrix $\mathbf{X}$ is denoted by $\mathsf{Diag}(\mathbf{x})$ or $\mathsf{Diag}(\mathbf{X})$, respectively.

We refer frequently to the following special matrices and sets: the $n \times n$ identity matrix is denoted $\mathbf{I}_n$, the $n \times n$ flipped identity matrix with ones on the anti-diagonal is denoted $\mathbf{J}_n$, the $m \times n$ matrix of all zeros is denoted $\mathbf{0}_{m\times n}$; the sub-scripts may be dropped where no confusion may arise. The set of all $n \times n$ unitary matrices is denoted $\mathbb{S}^n$ and the set of $m \times n$ complex matrices is denoted $\mathbb{C}^{m\times n}$.

The notation $\mathbf{x} \sim \mathcal{CN}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a random variable $\mathbf{x}$ which is circularly symmetric complex Gaussian distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

III. Modelling and Design Approach

In CS, we aim to reconstruct the signal of interest $\mathbf{x} \in \mathbb{C}^m$ based on a small number of noisy projections:

$$\mathbf{y} = \mathbf{M}\,\mathbf{x} + \mathbf{w}, \quad \mathbf{y} \in \mathbb{C}^\ell \tag{1}$$

with $\ell \leq m$ and where $\mathbf{M} \in \mathbb{C}^{\ell\times m}$ is the kernel (or projection) matrix and $\mathbf{w}$ represents zero-mean circularly symmetric complex Gaussian noise with positive definite covariance matrix $\boldsymbol{\Sigma}_{\mathbf{w}}$, i.e., $\mathbf{w} \sim \mathcal{CN}(\mathbf{w}; \mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{w}})$. The action of the kernel can be understood in terms of two separate projections and a power allocation (or stretching) operation, which are associated with the matrices in its singular value decomposition (SVD) given by:

$$\mathbf{M} = \mathbf{U}_{\mathbf{M}}\ \boldsymbol{\Lambda}_{\mathbf{M}}\ \mathbf{V}_{\mathbf{M}}^\dagger \tag{2}$$

where $\boldsymbol{\Lambda}_{\mathbf{M}} = \left[\mathsf{Diag}\left(\sqrt{\lambda_{M_1}}, \ldots, \sqrt{\lambda_{M_\ell}}\right) \quad \mathbf{0}_{\ell\times(m-\ell)}\right] \in \mathbb{R}^{\ell\times m}$, $\mathbf{U}_{\mathbf{M}} \in \mathbb{S}^\ell$, $\mathbf{V}_{\mathbf{M}} \in \mathbb{S}^m$, and $\lambda_{M1} \geq \lambda_{M2} \geq \ldots \geq \lambda_{M_\ell} \geq 0$ correspond to the (non-negative) eigenvalues of $\mathbf{M}\,\mathbf{M}^\dagger$.

Both the signal and the noise covariance matrices are positive (semi-)definite and can also be represented in terms of their eigenvalue decomposition (projections and power allocation). In particular, the signal covariance matrix is given by:

$$\boldsymbol{\Sigma}_{\mathbf{x}} = \mathbf{U}_{\mathbf{x}}\ \boldsymbol{\Lambda}_{\mathbf{x}}\ \mathbf{U}_{\mathbf{x}}^\dagger \tag{3}$$

where $\mathbf{U}_{\mathbf{x}} \in \mathbb{S}^m$, $\boldsymbol{\Lambda}_{\mathbf{x}} = \mathsf{Diag}\left(\lambda_{x_1}, \ldots, \lambda_{x_m}\right)$ and $\lambda_{x_1} \geq \lambda_{x_2} \geq \ldots \geq \lambda_{x_m} \geq 0$ are the (non-negative) eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{x}}$. Similarly, the noise covariance matrix is given by:

$$\boldsymbol{\Sigma}_{\mathbf{w}} = \mathbf{U}_{\mathbf{w}}\ \boldsymbol{\Lambda}_{\mathbf{w}}\ \mathbf{U}_{\mathbf{w}}^\dagger \tag{4}$$

where $\mathbf{U_w} \in \mathbb{S}^\ell$, $\mathbf{\Lambda_w} = \mathsf{Diag}\left(\lambda_{w_1}, \ldots, \lambda_{w_\ell}\right)$ and $0 \leq \lambda_{w1} \leq \lambda_{w2} \leq \ldots \leq \lambda_{w\ell}$ correspond to the (non-negative) eigenvalues of $\mathbf{\Sigma_w}$.

Our design approach, which relies not only on a statistical model for the noise but also on the signal, draws on specific quantitative metrics in order to conceive and compare various possible kernel designs. A natural metric, which relates to the best achievable reconstruction error, is the (non-linear) MMSE given by:

$$\mathrm{MMSE} = \mathbb{E}\left\{\mathsf{tr}\left[\left(\mathbf{x} - \mathbb{E}\left\{\mathbf{x}|\mathbf{y}\right\}\right)\left(\mathbf{x} - \mathbb{E}\left\{\mathbf{x}|\mathbf{y}\right\}\right)^\dagger\right]\right\} \tag{5}$$

that involves the use of conditional mean estimation to recover the signal of interest from the noisy projections, i.e., $\hat{\mathbf{x}}(\mathbf{y}) = \mathbb{E}\left\{\mathbf{x}|\mathbf{y}\right\}$ [34]. We, however, capitalize on information-theoretic metrics, most notably the mutual information and Rényi entropy based on the fact that mutual information and Rényi entropy - in view of recent developments in information theory and communications - appear to be more amenable to mathematical analysis than the non-linear MMSE. In addition, it is also possible to bound the MMSE via the mutual information as follows [30]:

$$\mathrm{MMSE} \geq \frac{1}{2\pi e} \exp 2\left[\mathcal{H}_x\left(\mathbf{x}\right) - \mathcal{I}\left(\mathbf{x};\mathbf{y}\right)\right]. \tag{6}$$

where $\mathcal{H}_x\left(\mathbf{x}\right)$ denotes the differential entropy of $\mathbf{x}$ and $\mathcal{I}\left(\mathbf{x};\mathbf{y}\right)$ denotes the mutual information between $\mathbf{x}$ and $\mathbf{y}$.

The crux of our design approach, which we also partially extend from the mutual information to the Rényi entropy metric, is a fundamental result that links the gradient with respect to some parameters of the mutual information between the input and the output of a linear vector Gaussian channel model and the MMSE matrix associated with the model: known as the I-MMSE relationship. This result, which was originally put forth for the linear scalar Gaussian model by Guo, Shamai and Verdú [27] and later for linear vector Gaussian channels by Palomar and Verdú in [28], can be directly applied to the model in (1) so that:

$$\nabla_{\mathbf{M}}\, \mathcal{I}\left(\mathbf{x};\mathbf{y}\right) = \mathbf{\Sigma_w}^{-1}\ \mathbf{M}\ \mathbf{E} \tag{7}$$

where the MMSE matrix is[3]:

$$\begin{aligned}\mathbf{E} &= \mathbb{E}\left\{\left(\mathbf{x} - \mathbb{E}\left\{\mathbf{x}|\mathbf{y}\right\}\right)\left(\mathbf{x} - \mathbb{E}\left\{\mathbf{x}|\mathbf{y}\right\}\right)^\dagger\right\} &(8)\\ &= \mathbf{U_E}\ \mathbf{\Lambda_E}\ \mathbf{U_E}^\dagger &(9)\end{aligned}$$

where $\mathbf{U_E} \in \mathbb{S}^m$, $\mathbf{\Lambda_E} = \mathsf{Diag}\left(\lambda_{E_1}, \ldots, \lambda_{E_m}\right)$ and $\lambda_{E1} \geq \lambda_{E2} \geq \ldots \geq \lambda_{Em} \geq 0$ are the (non-negative) eigenvalues of the MMSE matrix.

Our design approach also draws on a specific kernel design constraint. It is important to recall that in CS applications the kernel design is typically set to obey unit-norm row constrains or, instead, orthonormal constraints [35]. In contrast, in communications applications the kernel (or precoder) design obeys a power (trace) constraint, which states that on average the rows have unit-norm. A paper that does consider this power constraint for CS is the work by Schnitter [16], however, this is unusual in the CS field. We adopt this more general constraint, which, in addition to leading to solutions with higher mutual information or Rényi entropy, enables the formulation of the design framework which the unit-norm rows constraint does not. The exception to this is the special case of adaptive online design in Section VI-B, where each row of the kernel is designed sequentially such that the two constraints coincide.

IV. Mutual Information based Kernel Design

In this section we consider the characterization of the kernel that maximises the mutual information of the model in (1), subject to a power constraint, for multivariate Gaussian sources and general multivariate sources. The optimal kernel design for multivariate Gaussian sources also provides a rationale for other kernel designs in subsequent sections, most notably, the PDS method (we extend the work of [33]). The design problem can be posed abstractly as follows:

$$\begin{aligned}&\underset{\mathbf{M}}{\text{maximize}} && \mathcal{I}\left(\mathbf{x};\mathbf{M}\ \mathbf{x} + \mathbf{w}\right)\\ &\text{subject to} && \frac{1}{\ell}\ \mathsf{tr}\left(\mathbf{M}\mathbf{M}^\dagger\right) \leq 1\end{aligned} \tag{10}$$

It is important to remark that this optimization problem is non-convex in general. The use of the fundamental result in (7), in addition to enabling the full or partial characterization of the solution, also leads to efficient computational procedures. We restate next the characterizations of the optimal kernel designs for Gaussian sources (Theorem 1) and general sources (Theorem 2), which also appear in slightly different forms in [25] [36] [24], in a manner that emphasizes the operational significance for CS applications.

A. Multivariate Gaussian Input Source

The characterization of the optimal kernel design for a multivariate complex-valued Gaussian source leverages the well-known closed-form mutual information expression given by:

$$\mathcal{I}\left(\mathbf{x};\mathbf{y}\right) = \log\det\left(\mathbf{I}_m + \mathbf{M}^\dagger\mathbf{\Sigma_w}^{-1}\mathbf{M}\ \mathbf{\Sigma_x}\right). \tag{11}$$

This simple closed-form expression allows the use of simple matrix identities, rather than the gradient result in (7), to obtain the solution to (10). The case when $\mathbf{\Sigma_x} = \mathbf{I}$ is well-known from communications theory and was recently applied in the design of measurement kernels by Schniter [16]. However, the

[3] Note that the MMSE matrix $\mathbf{E}$ is a function of the kernel $\mathbf{M}$

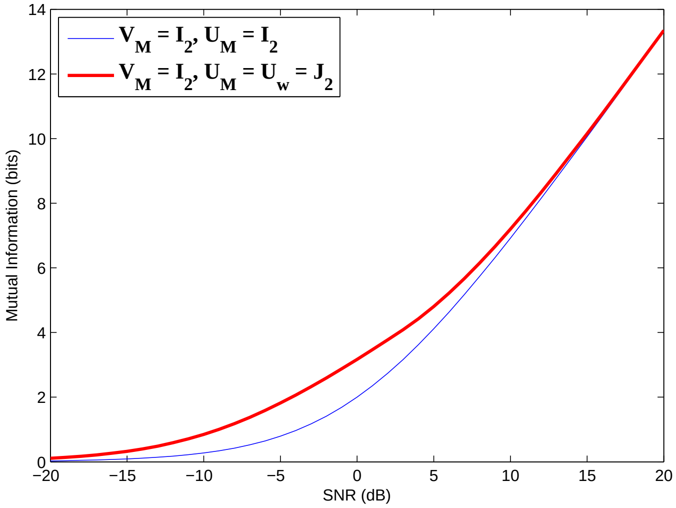
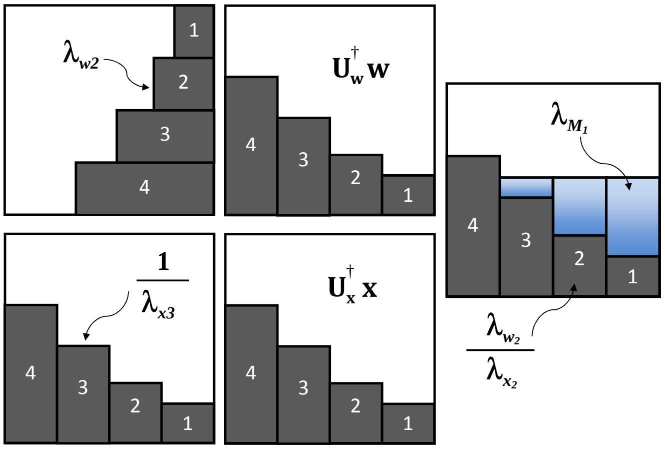


Fig. 1. Mutual information as a function of SNR for two different alignments both with optimal power allocation.



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

Fig. 2. Diagrammatic view of the actions of the optimal kernel design.

case for general source covariance matrices has not been studied in the communications domain. We unveil that this leads to the novel operation of *mode alignment*[4].

Theorem 1. *The kernel matrix that solves the optimization problem in* (10) *for a multivariate complex-valued Gaussian source with covariance matrix* $\boldsymbol{\Sigma}_{\mathbf{x}}$ *is given by*[5]*:*

$$\mathbf{M}^{\star} = \mathbf{U}_{\mathbf{w}}\ \boldsymbol{\Lambda}_{\mathbf{M}}^{\star}\ \mathbf{U}_{\mathbf{x}}^{\dagger} \tag{12}$$

where $\boldsymbol{\Lambda}_{\mathbf{M}}^{\star} = \left[\mathsf{Diag}\left(\sqrt{\lambda_{M_1}^{\star}}, \ldots, \sqrt{\lambda_{M_\ell}^{\star}}\right) \quad \mathbf{0}_{\ell\times(m-\ell)}\right]$, $\lambda_{M_i}^{\star} = \left(\frac{1}{\eta} - \frac{\lambda_{w_i}}{\lambda_{x_i}}\right)^{+}$ *with the noise covariance eigenvalues* λ_{wi} *arranged in ascending order and the source covariance eigenvalues* λ_{xi} *arranged in descending order and* η *ensures the* average *unit-norm row constraint, i.e.,* $\frac{1}{\ell}\ \mathsf{tr}\left(\mathbf{M}\mathbf{M}^{\dagger}\right) = 1$.

Proof: See Appendix C. ■

Theorem 1 uncovers the operations of the optimal kernel design. In particular, it is possible to recognize a novel mode alignment operation which involves two aspects: i) exposing the modes of the noise and source covariance; and ii) ordering (or aligning) the modes.

First, the left-singular vectors of the kernel are chosen to align with the eigenvectors of the noise covariance matrix and the right-singular vectors of the kernel are chosen to align with the eigenvectors of the signal covariance matrix (Fig. 2(a)). This is referred to as exposing the modes.

The ordering (or alignment) of the exposed modes is very particular, the largest source eigenvalue is matched to the smallest noise eigenvalue, the second largest source eigenvalue is matched to the second smallest noise eigenvalue, and so on (Fig. 2(b)).

Finally, the kernel "weights" the matched modes according to a "waterfilling" interpretation [37] (Fig. 2(c)). Intuitively, this emphasizes the less noisy "channels" and reduces the influence of the noisier ones as a means to maximize further mutual information.

As an example, Fig. 1 depicts the mutual information associated with two possible alignments for the signal and noise eigenvalues in a scenario where both covariance matrices are diagonal, $\boldsymbol{\Lambda}_{\mathbf{x}} = \mathsf{Diag}\,(1, 0.25)$ and $\boldsymbol{\Lambda}_{\mathbf{w}} = \mathsf{Diag}\,(1, 0.25)$. It is evident that the ordering of the modes has a significant impact on the mutual information at low and medium SNR – the highest mutual information corresponds to the kernel design that aligns the strongest source eigenvalue with the weakest noise eigenvalue, $\mathbf{U}_{\mathbf{M}} = \mathbf{U}_{\mathbf{w}} = \mathbf{J}_2$ and $\mathbf{V}_{\mathbf{M}} = \mathbf{I}_2$.

B. General Multivariate Input Source

While the application of communications theory results for Gaussian distributions are known to varying degrees outside the field of communications theory, the results for general sources have not been fully leveraged outside of communications. The characterization of the optimal kernel design for a general multivariate complex-valued source, in view of the absence of closed-form mutual information expressions, now leverages the fundamental result in (7).

Theorem 2. *The kernel matrix that solves the optimization problem in* (10) *for a general multivariate*

[4] This result was also recently shown in radar [32].

[5] Note that the superscript $\star$ denotes an optimal solution.

complex-valued source with covariance matrix $\mathbf{\Sigma}_{\mathbf{x}}$ is given by:

$$\mathbf{M}^\star = \mathbf{U}_{\mathbf{w}}\ \mathbf{\Lambda}^\star_{\mathbf{M}}\ \mathbf{V}^{\star\dagger}_{\mathbf{M}} \tag{13}$$

where $\mathbf{V}^\star_{\mathbf{M}} = \mathbf{U}^\star_{\mathbf{E}}\ \mathbf{\Pi}^\star$ [6]*, the matrix $\mathbf{\Pi}^\star$ is the optimal permutation matrix, $\mathbf{\Lambda}^\star_{\mathbf{M}} = \left[\mathsf{Diag}\left(\sqrt{\lambda^\star_{M_1}}, \ldots, \sqrt{\lambda^\star_{M_\ell}}\right)\ \mathbf{0}\right]$, and $\lambda^\star_{M_i}$ are given by the generalized mercury waterfilling solution, i.e.,*

$$\lambda^\star_{M_i} = \begin{cases} 0, & \eta\ \lambda_{w_i} > \mathsf{mmse}_i\left(\mathbf{U}^\star_{\mathbf{E}}\ \mathbf{\Pi}^\star, \mathbf{\Lambda}^\star_{\mathbf{Q}}|_{\lambda^\star_{M_i}=0}\right) \\ \mathsf{mmse}_i^{-1}\left(\eta\ \lambda_{w_i}\right), & \textit{otherwise} \end{cases} \tag{14}$$

where η ensures the average *unit-norm row constraint, i.e., $\frac{1}{\ell}\ \mathsf{tr}(\mathbf{M}^\star\mathbf{M}^{\star\dagger}) = 1$, $\mathbf{\Lambda}^\star_{\mathbf{Q}} = \mathbf{\Lambda}^\star_{\mathbf{M}}\mathbf{\Lambda}^{\star\dagger}_{\mathbf{M}}$, $\mathbf{\Lambda}^\star_{\mathbf{Q}}|_{\lambda^\star_{M_i}=0} = \mathsf{Diag}\left(\lambda^\star_{M_1}, \ldots, \lambda^\star_{M_{i-1}}, 0, \lambda^\star_{M_{i+1}}, \ldots, \lambda^\star_{M_\ell}\right)$ and $\mathsf{mmse}_i\left(\mathbf{U}^\star_{\mathbf{E}}\ \mathbf{\Pi}^\star, \mathbf{\Lambda}^\star_{\mathbf{Q}}\right)$ denotes the i-th diagonal entry of the MMSE matrix associated with the estimate of $\mathbf{x}' = \mathbf{\Pi}^{\star\dagger}\mathbf{U}^{\star\dagger}_{\mathbf{E}}\ \mathbf{x}$ from*

$$\mathbf{y}' = \mathbf{\Lambda}_{\mathbf{w}}^{-1/2}\ \mathbf{\Lambda}^\star_{\mathbf{M}}\ \mathbf{x}' + \mathbf{n}, \tag{15}$$

where $\mathbf{n}$ is zero-mean circularly symmetric complex Gaussian noise with identity covariance, $\mathbf{n} \sim \mathcal{CN}(\mathbf{n}; \mathbf{0}, \mathbf{I})$. Note that mmse_i^{-1} is the inverse of mmse_i with respect to λ_{Mi} for fixed $\lambda_{M_j}, \forall j \neq i$.

Proof: See Appendix D. ■

Remark 1. *In the high noise/low signal power regime, a first-order expansion of the mutual information is given [28]:*

$$\mathcal{I}(\mathbf{x}; \mathbf{y}) = \frac{1}{2}\ \mathsf{tr}\left(\mathbf{\Sigma}_{\mathbf{w}}^{-1}\ \mathbf{M}\ \mathbf{\Sigma}_{\mathbf{x}}\ \mathbf{M}^\dagger\right) + o(||\mathbf{\Sigma}_{\mathbf{x}}||) \tag{16}$$

which implies the result observed by Shannon that at low signal-to-noise ratios proper complex discrete inputs offer a negligible loss in performance terms with regards to the capacity achieved by Gaussian inputs; hence the results for the Gaussian in Theorem 1 also apply in general for proper complex sources in the high noise/low power regime.

Theorem 2 suggests that the *mode alignment* is no longer between the eigenvectors of the source covariance and the eigenvectors of the noise covariance, but between the eigenvectors of the MMSE matrix and the eigenvectors of the noise covariance. The diagonalization of the MMSE matrix was first noted for communications by Lamarca [24] for identity source covariances, and the same holds true for CS for general source covariances. The singular values of the kernel are described by the mercury waterfilling algorithm [22] [25] which differs from waterfilling by adjusting for the non-Gaussian nature of the inputs, however, the procedure is remarkably similar.

It is important to emphasize that Theorem 1 characterizes fully the optimal kernel design but - in view of the non-convexity of the problem - Theorem 2 characterizes partially, via a fixed point equation, the optimal kernel since $\mathbf{U}^\star_{\mathbf{E}}$ is still a function of $\mathbf{M}$. The characterization is useful because it leads to i) stopping criteria for gradient descent algorithms via (7); and ii) alternative optimization algorithms. Note that if we implement gradient descent with (7) we may get trapped in local maxima since it is known that the mutual information is not always a concave function of $\mathbf{M}$ [23]. However, mutual information is known to be concave in the squared singular values of $\mathbf{M}$, for $\mathbf{U}_{\mathbf{M}} = \mathbf{U}_{\mathbf{w}}$ and fixed $\mathbf{V}_{\mathbf{M}}$. An alternative gradient descent algorithm that leads to the global maximum by avoiding local maxima switches between optimizing the singular values and the right-singular vectors of the kernel [26].

V. Design with Rényi Entropy

We consider the characterization of the kernel that maximizes the output Rényi entropy of the model in (1), subject to a power constraint, for multivariate Gaussian sources and multivariate Gaussian mixture sources. The design problem can then be posed as follows:

$$\begin{aligned} &\underset{\mathbf{M}}{\text{maximize}} && \mathrm{h}_\alpha\left(\mathbf{M}\ \mathbf{x} + \mathbf{w}\right) \\ &\text{subject to} && \frac{1}{\ell}\ \mathsf{tr}\left(\mathbf{M}\mathbf{M}^\dagger\right) \leq 1 \end{aligned} \tag{17}$$

where

$$\mathrm{h}_\alpha(\mathbf{y}) = \frac{1}{1-\alpha}\ \log \int p^\alpha(\mathbf{y}) d\mathbf{y}. \tag{18}$$

Note that Rényi entropy represents a generalization of Shannon entropy given by:

$$\mathrm{h}_s(\mathbf{y}) = -\int p(\mathbf{y}) \log p(\mathbf{y}) d\mathbf{y} \tag{19}$$

which is the special case when $\alpha = 1$.

A. Multivariate Gaussian Input Source

For multivariate Gaussian sources, both Shannon entropy and Rényi entropy can be expressed analytically for all values of $\alpha > 0$. In particular, the two are shown to be related in the following theorem[7]:

Theorem 3. *For a multivariate Gaussian input source where $\mathbf{x} \sim \mathcal{CN}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}_{\mathbf{x}})$, the Rényi entropy of order $\alpha > 0$ and the Shannon entropy associated with the output of the model in* (1) *are related as:*

$$h_\alpha(\mathbf{y}) = \ h_s(\mathbf{y}) - \ell\left(1 - \frac{\log \alpha}{\alpha - 1}\right), \tag{20}$$

where $\ h_s(\mathbf{y}) = \log\left[(2\pi e)^\ell \det\left(\mathbf{\Sigma}_{\mathbf{w}} + \mathbf{M}\mathbf{\Sigma}_{\mathbf{x}}\mathbf{M}^\dagger\right)\right].$

[6] Note that the MMSE matrix $\mathbf{E}$ is a function of the kernel $\mathbf{M}$.

[7] For quadratic Rényi entropy this result was also in Appendix A of [38].

Proof: See Appendix E. ■

Theorem 3 leads immediately to a generalization of the I-MMSE identity in (7) for Gaussian sources:

Theorem 4. *For Gaussian sources, the (complex) gradient with respect to the kernel of the output Rényi entropy of order $\alpha > 0$ associated with the model in* (1) *obeys the relationship:*

$$\nabla_{\mathbf{M}}\ h_\alpha(\mathbf{y}) = \boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\ \mathbf{M}\ \mathbf{E}. \tag{21}$$

Theorem 4 unveils that the relationship between mutual information and the MMSE matrix in (7) also holds for all values of $\alpha > 0$ for the output Rényi entropy associated with the model in (1) for Gaussian sources. Theorem 4 also implies that the kernel design that maximizes the Rényi entropy subject to a power constraint also obeys the characterization in Theorem 1.

B. Multivariate Gaussian Mixture Model Input Source

For Gaussian Mixture Models (GMM) the signal $\mathbf{x} \in \mathbb{C}^m$ is represented by:

$$p(\mathbf{x}) = \sum_{i=1}^{N} p(i)\ \mathcal{CN}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{22}$$

where $p(i)$ is the probability of occurrence of mixture component i, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ correspond to the mean and covariance matrix of the i-th circularly symmetric complex Gaussian distribution. Neither the Shannon entropy, mutual information nor the MMSE matrix are known to have closed-form expressions for GMMs. Rényi entropy and its gradient, however, admit closed-form expressions in some instances, which lend themselves more easily to optimization via gradient descent algorithms. For example, the quadratic Rényi entropy of the noisy projection $\mathbf{y}$ in (1) is given by:

$$\mathrm{h}_2(\mathbf{y}) = -\log \sum_{i=1}^{N}\sum_{j=1}^{N} p(i)\ p(j)\ \mathcal{CN}\left(\mathbf{0}; \boldsymbol{\mu}_{i,j}, \boldsymbol{\Sigma}_{i,j}\right) \tag{23}$$

where:

$$\boldsymbol{\mu}_{i,j} = \mathbf{M}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) \tag{24}$$

$$\boldsymbol{\Sigma}_{i,j} = \mathbf{M}\left(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j\right)\mathbf{M}^\dagger + 2\boldsymbol{\Sigma}_{\mathbf{w}} \tag{25}$$

The complex gradient with respect to M of the quadratic Rényi entropy of the noisy projection $\mathbf{y}$ in (1) for the GMM is given by:

$$\nabla_{\mathbf{M}}\ \mathrm{h}_2(\mathbf{y}) = \frac{-\sum_{i,j=1}^{N} p(i)\ p(j)\ \mathcal{CN}\left(\mathbf{0}; \boldsymbol{\mu}_{i,j}, \boldsymbol{\Sigma}_{i,j}\right)\nabla_{\mathbf{M}}\log\mathcal{CN}\left(\mathbf{0}; \boldsymbol{\mu}_{i,j}, \boldsymbol{\Sigma}_{i,j}\right)}{\sum_{i=1}^{N}\sum_{j=1}^{N} p(i)\ p(j)\ \mathcal{CN}\left(\mathbf{0}; \boldsymbol{\mu}_{i,j}, \boldsymbol{\Sigma}_{i,j}\right)} \tag{26}$$

where:

$$\begin{aligned}\nabla_{\mathbf{M}}\log\mathcal{CN}_{i,j} = &-\boldsymbol{\Sigma}_{i,j}^{-1}\ \mathbf{M}\ \left(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j\right)\\ &+\boldsymbol{\Sigma}_{i,j}^{-1}\ \mathbf{M}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^\dagger\\ &\times\left\{\mathbf{M}^\dagger\boldsymbol{\Sigma}_{i,j}^{-1}\mathbf{M}\left(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j\right) - \mathbf{I}\right\}.\end{aligned} \tag{27}$$

where × denotes a matrix multiplication. The proof is given in Appendix F.

It is interesting to note that the now celebrated I-MMSE relationship in the information theory literature also applies for Rényi entropy of order $\alpha > 0$ associated with Gaussian source models. However, this relationship does not seem to carry over for the Rényi entropy of more general source models. In fact, it can only be shown that for a general source, which obeys some additional smoothness conditions, the gradient can be expressed as follows (The proof is a modification of the result in [28]):

$$\nabla_{\mathbf{M}}\ \mathrm{h}_\alpha(\mathbf{y}) = \alpha\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\int \hat{p}(\mathbf{y})\left(\mathbf{y} - \mathbf{M}\mathbf{x}_y\right)\mathbf{x}_y^\dagger d\mathbf{y} \tag{28}$$

where the probability distribution $\hat{p}(\mathbf{y}) = \frac{p^\alpha(\mathbf{y})}{\int p^\alpha(\mathbf{y})d\mathbf{y}}$ and $\mathbf{x}_y$ is the conditional mean estimator.

It is not difficult to appreciate that the right-hand side of (28) is in general different from the right-hand side of (21) (or the right-hand side of the I-MMSE relationship in (7)) by studying the Taylor expansion of $\nabla_{\mathbf{M}}\ \mathrm{h}_2(\mathbf{y})$. For the high-noise power scenario, the first term in the expressions coincide but higher order terms do not [39].

VI. Application to Compressive Sensing

A. Problem setup

We consider CS in the context of imaging. While the theory is applicable to complex data, the following examples focus on real images. Specifically, consider measurement of the image $\mathbf{X} \in \mathbb{R}^{N_x \times N_y}$, for large N_x and N_y. As indicated in Figure 3, the image is partitioned into $n_x \times n_y$ contiguous "patches," with the pixels in the jth patch denoted by vector $\mathbf{x}_j \in \mathbb{R}^\ell$, with $\ell = n_x n_y$. In the examples considered here $n_x = n_y = 8$ (consistent, for example, with the patch sizes used in the JPEG standard).

It is desirable to partition the images into such patches because one may readily learn a signal model for the $\{\mathbf{x}_j\}$, while it is difficult to learn an accurate signal model directly on the entire image $\mathbf{X}$. Specifically, following [13], [33], we assume that each $\mathbf{x}_j$ is drawn from a GMM of the form (22), here for *real* normal distributions.

To learn the prior signal model $p(\mathbf{x})$ for the patches, we first consider a large ensemble of natural images, from which patches $\mathbf{x}_j \in \mathbb{R}^\ell$ are selected at random. Using these training data, a (real) GMM of the form in (22) is constituted as a signal model. To learn this GMM, we have employed nonparametric Bayesian methods as in [13], as well as expectation-maximization (EM) methods [33], and both methods yield very similar results. The following results are based on a $N = 20$ component GMM, trained on 100,000 patches, extracted at random from 500 natural images in the Berkeley Segmentation Dataset (http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/resources.html). These training images are distinct from those considered in the testing phase, for CS inversion.

While patches are selected at random from training images to constitute the prior $p(\mathbf{x})$, when performing CS the goal is to recover the entire underlying image $\mathbf{X}$. Therefore, for CS inversion we wish to recover each of the $\{\mathbf{x}_j\}$ in Figure 3. In general, a separate projection matrix $\mathbf{M}_j$ is applied to patch j from image $\mathbf{X}$. For the case of offline design of the projection matrix, $\mathbf{M}_j$ is the same for all patches j (since it is non-adaptive). For online design a distinct $\mathbf{M}_j$ is adaptively designed for each testing patch j. The measured data associated with patch j is expressed as

$$\mathbf{y}_j = \mathbf{M}_j\mathbf{x}_j + \mathbf{w}_j, \ j = 1, \ldots, J \tag{29}$$

In the examples that follow, the images under test are 256×256, and therefore this procedure was employed on $J = 1024$ non-overlapping patches of size 8×8. Each of the $\mathbf{x}_j$ are recovered independently from the respective measured $\mathbf{y}_j$, thereby allowing for massive parallelization.

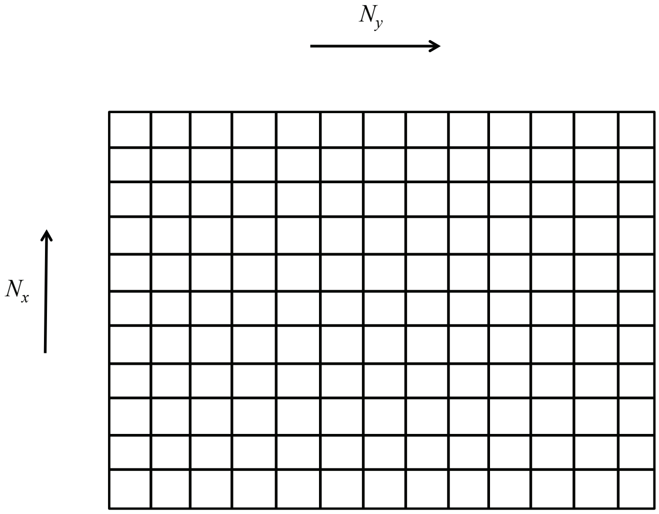


Fig. 3. Spatial grid used for CS measurement image $N_x \times N_y$ image, decomposing the image into a contiguous grid of "patches," each patch composed of $n_x \times n_y$ pixels, $n_x \ll N_x$ and $n_y \ll N_y$. Letting $\mathbf{x}_i \in \mathbb{R}^{n_x n_y}$ represent the pixels associated with the ith patch, separate projection matrices $\mathbf{M}_i$ are designed for each $\mathbf{x}_i$.

For simplicity, we henceforth drop the subscript j, and the discussion that follows applies to each of the J patches in Figure 3. We assume the noise $\mathbf{w} \sim \mathcal{N}(0, \mathbf{\Sigma}_\mathbf{w})$, with known covariance matrix $\mathbf{\Sigma}_\mathbf{w}$. In the following examples we consider low-noise, i.i.d. measurements, and therefore $\mathbf{\Sigma}_\mathbf{w} = 10^{-6}\mathbf{I}_\ell$. The likelihood function for the underlying signal $\mathbf{x}$ is $\mathcal{N}(\boldsymbol{y}; \mathbf{\Phi}\boldsymbol{x}, \mathbf{\Sigma}_\mathbf{w})$, and the prior $p(\mathbf{x})$ is the aforementioned GMM, $p(\boldsymbol{x}) = \sum_{i=1}^N w_i\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$. Under this likelihood function for $\mathbf{x}$, and with the GMM prior, the posterior $p(\mathbf{x}|\mathbf{y})$ is also a GMM:

$$p(\boldsymbol{x}|\boldsymbol{y}) = \sum_{i=1}^N \widetilde{w}_i\mathcal{N}(\boldsymbol{x}; \widetilde{\boldsymbol{\mu}}_i, \widetilde{\mathbf{\Sigma}}_i) \tag{30}$$

with

$$\widetilde{\mathbf{\Sigma}}_i^{-1} = \mathbf{M}^T\mathbf{\Sigma}_\mathbf{w}^{-1}\mathbf{M} + \mathbf{\Sigma}_i^{-1}, \ \widetilde{\boldsymbol{\mu}}_i = \widetilde{\mathbf{\Sigma}}_i(\mathbf{M}^T\mathbf{\Sigma}_\mathbf{w}^{-1}\boldsymbol{y} + \mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i) \tag{31}$$

$$\widetilde{w}_i = w_i\mathcal{N}(\boldsymbol{y}; \mathbf{M}\boldsymbol{\mu}_i, \mathbf{M}\mathbf{\Sigma}_i\mathbf{M}^T + \mathbf{\Sigma}_\mathbf{w})/p(\boldsymbol{y}) \tag{32}$$

$$p(\boldsymbol{y}) = \sum_{i=1}^N w_i\mathcal{N}(\boldsymbol{y}; \mathbf{M}\boldsymbol{\mu}_i, \mathbf{M}\mathbf{\Sigma}_i\mathbf{M}^T + \mathbf{\Sigma}_\mathbf{w}) \tag{33}$$

When presenting results, the estimated signal $\hat{\mathbf{x}}$ is the mean based on $p(\mathbf{x}|\mathbf{y})$, *i.e.*, $\hat{\mathbf{x}} = \sum_{i=1}^N \widetilde{w}_i\widetilde{\boldsymbol{\mu}}_i$.

B. Offline and online design

We consider online and offline design of the projection matrix $\mathbf{M}$, based upon gradient descent: $\mathbf{M} \leftarrow \mathbf{M} + \gamma\nabla_\mathbf{M}\mathcal{I}(\boldsymbol{x}; \boldsymbol{y})$, with re-normalization to satisfy the power constraint; here we perform a gradient of the mutual information, and the same type of gradient descent is performed in the context of Rényi entropy, for which we therefore employ the results of Section IV. When employing the gradient of Rényi entropy, we employ (7). The design of $\mathbf{M}$ based upon a gradient of mutual information is denoted PV, for Palomar and Verú.

For offline PV design, $p(\mathbf{x})$ corresponds to the learned prior GMM, and the entire $\mathbf{M}$ is inferred at once. For online PV design, after measuring the first k components of $\mathbf{y}$, denoted $\mathbf{y}_{1:k}$, we update the posterior $p(\mathbf{x}|\mathbf{y}_{1:k})$ via (30), and row $k+1$ of $\mathbf{M}$ is constituted based upon this posterior signal model; after each measurement, the posterior is updated, followed by design of the next row of $\mathbf{M}$, used to define the next measurement. In these computations, the MMSE matrix in (8) is computed via Monte Carlo integration, based on draws from $p(\mathbf{x})$ (in the offline case) or $p(\mathbf{x}|\mathbf{y}_{1:k})$ (in the online case). Online design of the patch-dependent projection matrix $\mathbf{M}$ may be performed in parallel.

For Rényi-based design we consider the case $\alpha = 2$; this is convenient, as within the context of the GMM representation employed here the gradient with respect to $\mathbf{M}$ is analytic, via (26).

The online PV design is relatively expensive, as one must repeatedly perform Monte Carlo integration to update the MMSE matrix $\mathbf{E}$, and one must also perform gradient descent. For online Rényi-based design we employ (26); while this analytic expression precludes the need to numerically compute $\mathbf{E}$, the large number of sums makes online Rényi and online PV design comparably expensive.

The relative expense of Rényi and PV online design motivates a simplified online design. In [33] the authors proposed the PDS method, in which a GMM was used for $p(\mathbf{x})$. In [33], the components of the first $k < \ell$ rows of M are drawn i.i.d. from a zero-mean normal distribution. Using this k-row sensing matrix, an initial measurement $\mathbf{y}_{1:k} \in \mathbb{R}^k$ is performed. Based upon $\mathbf{y}_{1:k}$, the most probable mixture component from the prior $p(\mathbf{x})$ is selected. At this point a single-Gaussian signal model is constituted. The remaining $\ell - k$ rows of M are then defined by the principal $\ell - k$ eigenvectors of the covariance matrix from this Gaussian. While [33] did not have access to our Theorem 1, the design so constituted is consistent with it. Specifically, Theorem 1 applies to the case of a single-Gaussian signal model. Under the aforementioned assumptions for $\mathbf{\Sigma}_\mathbf{w}$ (diagonal covariance matrix, with small diagonal variance), Theorem 1 implies that the optimal projection matrix corresponds to the principal eigenvectors of the covariance matrix. However, the assumption of k initial random projections employed in [33], before selecting a single Gaussian component, seems undesirable. Further, in [33] the single Gaussian was selected from the prior $p(\mathbf{x})$ rather from the updated posterior $p(\mathbf{x}|\mathbf{y}_{1:k})$.

We extend the PDS technique to an online setting as follows. We first initialize $p(\boldsymbol{x})$ with the GMM prior signal model (learned using offline training data). We then sequentially constitute one row of M at a time, from $k = 1, \ldots, \ell$; after each row is so constituted, a single new projection measurement is performed with that new row. Again let $\mathbf{y}_{1:k}$ represent the vector of data constituted in this manner via the first k rows of M. Based upon these data we update the signal model $p(\mathbf{x}|\mathbf{y}_{1:k})$. To design row $k+1$ of M, let $i' = \operatorname{argmax}_i \widetilde{w}_i$, where the $\{\widetilde{w}_i\}$ are the GMM mixture weights from $p(\mathbf{x}|\mathbf{y}_{1:k})$. Then the $(k+1)$th row of M is defined by the leading eigenvector of $\widetilde{\mathbf{\Sigma}}_{i'}$. The online PDS approximates the posterior GMM at each step with the dominant Gaussian from the posterior GMM $p(\mathbf{x}|\mathbf{y}_{1:k})$, and then via Theorem 1 the next row of M is defined by the leading eigenvector of the associated covariance matrix. Since no Monte Carlo simulation and gradient descent are needed in the above process, online PDS method is very fast. The eigenvectors are orthonormal, and therefore the power constraint is satisfied automatically at every step. Note that the posterior $p(\mathbf{x}|\mathbf{y}_{1:k})$ continuously updates with increasing data, and therefore it is not particularly sensitive to the prior $p(\mathbf{x})$; the original PDS in [33] was based upon the prior $p(\mathbf{x})$ only, which may necessitate more care in selection of the training data. Since the posterior can be updated easily via (30), it appears highly preferable to use this approach rather than fixing the signal model.

C. Experimental Results

In Figures 4-6 results are shown for three widely examined test images: 'barbara', 'house' and 'pepper,' respectively. Two classes of results are considered based upon random projection design. The "random GMM" results employ the patch-based CS construction in Figure 3, and the learned GMM-based prior $p(\mathbf{x})$. The form of these results are the same as employed for the designed $\mathbf{M}_j$, except here each $\mathbf{M}_j$ is constituted with matrix elements drawn i.i.d. from $\mathcal{N}(0, 1)$, followed by normalization. We also considered CS design in which the projections are performed directly on the entire image $\mathbf{X}$, rather than at the patch level, as in Figure 3. If one performs CS inversion based on traditional CS algorithms, which employ ℓ_1 and related regularization [40], the quality of the inversion is markedly worse than that using the proposed approach, with learned signal models $p(\mathbf{x})$; we therefore do not show these results here, because they don't fit on the same scale of results presented. This is not surprising, as the patch-dependent learned signal model $p(\mathbf{x})$ is much richer, and tailored to the data than simple sparsity constraints, which motivate ℓ_1 regularization. To provide a fairer comparison, when performing inversion for the case in which the projections are performed directly on the entire image $\mathbf{X}$, we consider an underlying wavelet basis and perform inversion based on the sophisticated hidden Markov tree (HMT) wavelet model for images [41]. This signal model $p(\mathbf{x})$ could in principle also be used within the theory to design a projection matrix applicable to the entire image. However, the significant advantage of the GMM construction is that the posterior of the underlying signal may be constituted analytically, while for the HMT expensive computational methods are needed [41]. Therefore, we only show HMT inversion results when the projection matrix is constituted at random, thereby providing a comparison of inversion quality of the GMM (patch based) and the HMT (entire image), based upon random projections.

We consider offline design of the patch-based projection matrix M based upon the Rényi measure of entropy, as well as based upon mutual information (via the PV theory). For online Rényi and PV design, we do *not* make a simplifying single-Gaussian assumption when designing each row of M. By contrast the online PDS method uses the most probable Gaussian from the posterior to design the next projection at each step (this is therefore an approximation). The PDS method is very fast, while online PV is expensive, and therefore is shown principally for comparison (may not be done in practice, where online design must be fast).

First comparing the results based on random projections, the results based upon the (learned)

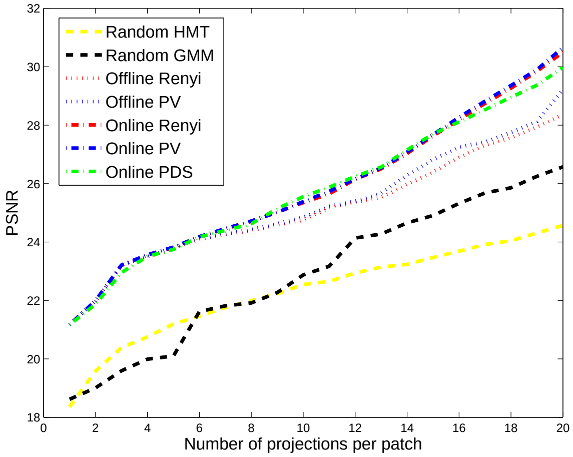


Fig. 4. PSNR for the reconstructed ‘barbara’ image.

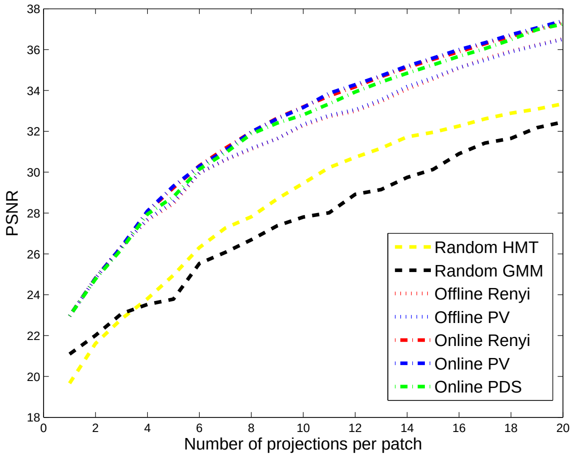


Fig. 5. PSNR for the reconstructed ‘house’ image.

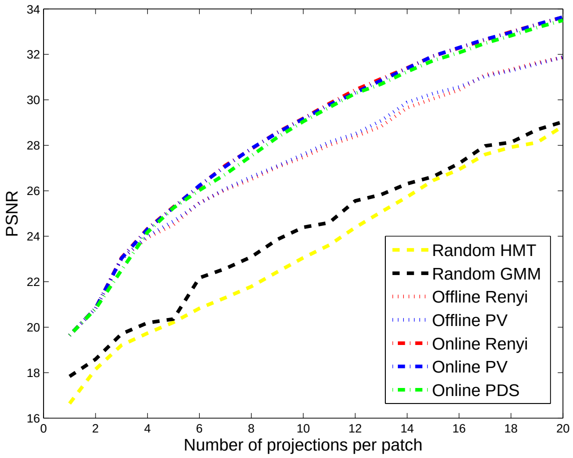


Fig. 6. PSNR for the reconstructed ‘pepper’ image.

patched-based GMM and based on the entire-image-based HMT are comparable in reconstruction quality. Sometimes the GMM results are slightly better, and other times the HMT results are better. However, there is no comparison with respect to computation speed. The HMT results are expensive, being based upon a Gibbs sampler [41]. By contrast the GMM results are very fast, with the inversion analytic. The additional big advantage of the GMM representation is that it allows convenient design of patch-dependent projection matrices, which we consider next.

Each of the designed projection methods yield significant improvement relative to random, and after approximately 6 projections per patch we note that the online results are significantly better than offline design. For the first approximately 5 measurements per patch, the offline and online results are comparable; we attribute this to an inadequate number of measurements to obtain an accurate signal model, and therefore little gain manifested by adaptivity. However, after approximately 6 measurements per patch it appears that the posterior signal model becomes accurate, yielding advantages of adaptivity. Concerning online design, inversion quality based on the simple and fast online PDS performs quite competitively relative to the online Rényi and PV design (which do not make a simplification to a single Gaussian), despite the fact that it assumes that the patch is drawn from a single Gaussian.

To understand the quality of the simple PDS-based design, consider Figure 7, wherein we plot the probabilities $\{\widetilde{w}_i\}_{i=1,N}$, for the posterior $p(\mathbf{x}|\mathbf{y}_{1:k})$, as the number of measurements k increases from 1 to ℓ. Note that after approximately six measurements the model has inferred that the underlying signal $\mathbf{x}$ was drawn from a single multivariate Gaussian. Note that the GMM is characteristic of an *ensemble* of draws, like those characteristic of the multiple patches in Figure 3. However, any single patch is drawn from a single one of the mixture components; it is however unknown *a priori* which component. Based upon experiments of this type, typically 6 projections are sufficient to infer which single mixture component a given patch corresponds to. At this point the results in Theorem 1 apply directly, which under the assumption for $\mathbf{\Sigma_w}$ dictates that the optimal measurement corresponds to projecting onto the dominant eigenvector of the covariance matrix of the single mixture component (single Gaussian); this is precisely what PDS does.

VII. Conclusions

We observe that the design principle of maximizing mutual information or Rényi entropy leads to deterministic kernel matrices for which MMSE performance is superior to that of random kernel

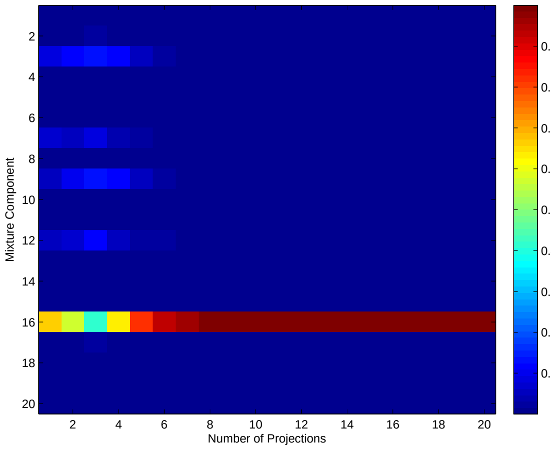


Fig. 7. Evolution of the mixture weight in the posterior GMM for a typical testing patch in 'barbara'.

matrices. In particular, we are able to provide design principles for the optimal kernel matrix for a general multivariate source that maximizes the mutual information or Rényi entropy (for which Shannon entropy is a special case). We showed that the optimal kernel 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 theoretical framework has been demonstrated with application to compressive sensing (CS) as applied to imagery. Using a GMM signal model, it was demonstrated that designed measurement kernels can yield markedly improved CS signal recovery relative to random design. The GMM representation has the advantage of yielding closed-form CS inversion, which is particularly attractive for fast signal inversion and for online kernel design. We have enhanced an online kernel design framework first proposed in [33], and have also provided a theoretical foundation for why it works so effectively in practice.

References

[1] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, pp. 1207–1223, 2006.

[2] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289 –1306, 2006.

[3] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing MRI," in *IEEE Signal Processing Magazine*, 2007.

[4] M. E. Gehm, R. John, D. J. Brady, R. M. Willett, and T. J. Schulz, "Single-shot compressive spectral imaging with a dual-disperser architecture," *Opt. Express*, vol. 15, pp. 14 013–14 027, 2007.

[5] T. Sun, C. Li, Y. Zhang, L. Xu, and K. Kelly, "Compressive hyperspectral acquisition and endmember unmixing," *Proc. SPIE 8165*, vol. 81650D, 2011.

[6] M. Shankar, N. Pitsianis, and D. Brady, "Compressive video sensors using multichannel imagers," *Appl. Opt.*, vol. 49, pp. B9–B17, 2010.

[7] Y. Hitomi, J. Gu, M. Gupta, T. Mitsunaga, and S. K. Nayar, "Video from a single coded exposure photograph using a learned over-complete dictionary," in *IEEE International Conference on Computer Vision (ICCV)*, Nov 2011.

[8] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, pp. 489–509, 2006.

[9] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Processing*, vol. 56, pp. 2346–2356, 2008.

[10] A. Ashok, P. Baheti, and M. Neifeld, "Compressive imaging system design using task-specific information," *Applied Optics*, vol. 47, no. 25, pp. 4457–4471, 2008.

[11] P. Baheti and M. Neifeld, "Recognition using information-optimal adaptive feature-specific imaging," *Journal of the Optical Society of America A*, vol. 26, no. 4, pp. 1055–1070, 2009.

[12] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *Information Theory, IEEE Transactions on*, vol. 56, no. 4, pp. 1982 –2001, 2010.

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

[14] Y. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Processing*, vol. 58, pp. 3042–3054, 2010.

[15] R. Baraniuk and M. Wakin, "Random projections of smooth manifolds," *Foundations of Computational Mathematics*, vol. 9, pp. 51–77, 2009.

[16] P. Schniter, "Exploiting structured sparsity in bayesian experimental design," in *CAMSAP '11*, 2011.

[17] W. R. Carson, M. R. D. Rodrigues, M. Chen, L. Carin, and R. Calderbank, "How to focus the discriminative power of a dictionary," in *ICASSP '12*, 2012.

[18] D. P. Palomar, J. Cioffi, and M. A. Lagunas, "Joint Tx-Rx beamforming design for multcarrier MIMO channels: A unified framework for convex optimization," in *IEEE Trans. Signal Process.*, Sept. 2003, pp. 2381–2401.

[19] A. Scaglione, P. Stoica, S. Barbarossa, G. B. Giannakis, and H. Sampath, "Optimal designs for space-time linear precoders and decoders," *IEEE Trans. Signal Processing*, vol. 50, pp. 1051–1064, 2002.

[20] S. Bergman and B. Ottersten, "Lattice-based linear precoding for MIMO channels with transmitter CSI," in *IEEE Trans. Signal Process.*, July 2008, pp. 2902–2914.

[21] S. Goparaju, A. R. Calderbank, W. R. Carson, M. R. Rodrigues, and F. Pérez-Cruz, "When to add another dimension when communicating over MIMO channels," in *accepted to ICASSP '11*, May 2011.

[22] 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.

[23] M. Payaró and D. P. Palomar, "Hessian and concavity of mutual information, entropy, and entropy power in linear vector Gaussian channels," in *IEEE Trans. Inf. Theory*, Aug. 2009, pp. 3613–3628.

[24] M. Lamarca, "Linear precoding for mutual information maximization in MIMO systems," in *Intl. Symposium on Wireless Communication Systems (ISWCS '09)*, Sept. 2009, pp. 26–30.

[25] 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.

[26] C. Xiao, Y. R. Zheng, and Z. Ding, "Globally optimal linear precoders for finite alphabet signals over complex vector gaussian channels," in *IEEE Trans. Signal Process.*, July 2011, pp. 3301–3314.

[27] 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.

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

[29] M. Hellman and J. Raviv, "Probability of error, equivocation, and the Chernoff bound," *IEEE Trans. Information Theory*, vol. 16, pp. 368–372, 1970.

[30] S. Prasad, "http://arxiv.org/pdf/1010.1508v1.pdf."

[31] D. Erdogmus and J. C. Principe, "Lower and upper bounds for misclassification probability based on Renyi's information," *J. VLSI Signal Proc.*, vol. 37, pp. 305–317, 2004.

[32] B. Tang, J. Tang, and Y. Peng, "MIMO radar waveform design in colored noise based on information theory," in *IEEE Trans. Signal Process.*, Sept. 2010, pp. 4684–4697.

[33] J. Duarte-Carvajalino, G. Yu, L. Carin, and G. Sapiro, "Adapted statistical compressive sensing: Learning to sense gaussian mixture models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012.

[34] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, 2001.

[35] Z. Nenadic, "Information discriminant analysis: Feature extraction with an information-theoretic objective," in *IEEE Trans. Pattern Anal. Mach. Intell.*, Jan. 2007, pp. 1394–1407.

[36] 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.

[37] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, Aug. 1991.

[38] L. Scardovi, "Information based control for state and parameter based estimation," Ph.D. dissertation, Genova University, http://www.montefiore.ulg.ac.be/ scardovi/pub/phd.pdf, Mar. 2005.

[39] F. Renna, "private communication," 2012.

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

[41] L. He and L. Carin, "Exploiting structure in wavelet-based bayesian compressive sensing," *Signal Processing, IEEE Transactions on*, vol. 57, no. 9, pp. 3488 –3497, 2009.

[42] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2005.

[43] H. Richter, "Zur Abschätzung von Matrizennormen," in *Math. Nachr.*, 1958, pp. 178–187.

[44] C. Köse and R. D. Wesel, "Universal Space-Time trellis codes."

[45] C. M. Theobald, "An inequality for the trace of the product of two symmetric matrices," in *Math. Proc. Camb. Phil. Soc.*, Oct. 1974, pp. 265–267.

[46] H. S. Witsenhausen, "A determinant maximization problem occuring in the theory of data communication," in *SIAM J. Appl. Math.*, Nov. 1975, pp. 515–522.

[47] D. Palomar and Y. Jiang, "MIMO transceiver design via majorization theory," in *Foundations and Trends in Communications and Information Theory*, Nov. 2006.

[48] K. Petersen and M. Pedersen, "The matrix cookbook," in *www2.imm.dtu.dk/pubdb/p.php?3274*, 2008.

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

Appendix A
Complex Derivatives and Gradients

Throughout the paper we adopt the definition of the formal partial complex derivative of a real-valued scalar function f with respect to a complex-valued variable x given by [28] [26]:

$$\frac{\partial f}{\partial x^*} \triangleq \frac{1}{2}\left[\frac{\partial f}{\partial \mathrm{Re}(x)} + j\frac{\partial f}{\partial \mathrm{Im}(x)}\right] \tag{34}$$

The definition of the complex gradient of a real-valued function f with respect to a complex-valued matrix $\mathbf{X}$ is given by:

$$\nabla_{\mathbf{X}} f \triangleq \frac{\partial f}{\partial \mathbf{X}^*} \tag{35}$$

where $[\nabla_{\mathbf{X}} f]_{ij} = {}^{\partial f}/_{\partial [\mathbf{X}^*]_{ij}}$.

Appendix B
Helpful Lemmas

In the proofs of the Theorems stated in this paper we will find the following lemmas helpful

Lemma 1 (Sylvester's Determinant Theorem)**.** *We have a "cyclic" property of determinants for two matrices* $\mathbf{A} \in \mathcal{C}^{n\times m}$ *and* $\mathbf{B} \in \mathcal{C}^{m\times n}$*:*

$$\det\left(\mathbf{I}_n + \mathbf{A}\mathbf{B}\right) = \det\left(\mathbf{I}_m + \mathbf{B}\mathbf{A}\right). \tag{36}$$

In the following four lemmas we denote two Hermitian matrices by $\mathbf{\Sigma_A}, \mathbf{\Sigma_B} \in \mathcal{C}^{m\times m}$ which have eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m$ and $\mu_1 \geq \cdots \geq \mu_m$, respectively. The eigenvalue decomposition of these two matrices are $\mathbf{\Sigma_A} = \mathbf{U_A}\ \mathbf{\Lambda_A}\ \mathbf{U_A^\dagger}$ and $\mathbf{\Sigma_B} = \mathbf{U_B}\ \mathbf{\Lambda_B}\ \mathbf{U_B^\dagger}$, where $\mathbf{U_A}, \mathbf{U_B} \in \mathbb{S}^{m\times m}$, $\mathbf{\Lambda_A} = \mathsf{Diag}\left(\lambda_1, \cdots, \lambda_m\right)$ and $\mathbf{\Lambda_A} = \mathsf{Diag}\left(\mu_1, \cdots, \mu_m\right)$.

Lemma 2 (Theorem 1.3.12 in [42])**.** *The matrices* $\mathbf{\Sigma_A}$ *and* $\mathbf{\Sigma_B}$ *commute if and only if they are simultaneously diagonalizable, i.e., both* $\mathbf{U}\mathbf{\Sigma_A}\mathbf{U}^\dagger$ *and* $\mathbf{U}\mathbf{\Sigma_B}\mathbf{U}^\dagger$ *are diagonal matrices for some unitary matrix* $\mathbf{U}$*.*

Lemma 3 (Richter [43])**.**

$$\sum_{i=1}^{m}\lambda_i\ \mu_{m+1-i} \leq \mathsf{tr}\left(\mathbf{\Sigma_A}\mathbf{\Sigma_B}\right) \leq \sum_{i=1}^{m}\lambda_i\ \mu_i. \tag{37}$$

Remark 2. *Sufficient conditions for achieving the upper and lower bounds are* $\mathbf{U_A} = \mathbf{U_B}$ *and* $\mathbf{U_A} = \mathbf{U_B}\ \mathbf{J}_m$*, respectively. The sufficient condition to achieve the lower bound was given by Köse and Wesel in Theorem 2 in [44] and Theobald [45] also gave necessary and sufficient conditions for achieving the upper bound, which allow for the multiplicity of eigenvalues.*

Lemma 4 (Lemma 3 in Witzenhausen [46])**.**

$$\det\left(\mathbf{I}_m + \mathbf{\Sigma_A}\mathbf{\Sigma_B}\right) \leq \prod_{i=1}^{m}\left(1 + \lambda_i\ \mu_i\right). \tag{38}$$

Remark 3. *A sufficient condition for achieving the upper bound is* $\mathbf{U_A} = \mathbf{U_B}$*. Witzenhausen gave further sufficient conditions which allow for the multiplicity of eigenvalues, stating that if equality holds then* $\mathbf{\Sigma_A}$ *and* $\mathbf{\Sigma_B}$ *commute and the diagonalizing matrix is such that the eigenvalues are aligned in the same order.*

Lemma 5. *Let* $\mathbf{P} \in \mathcal{C}^{m\times n}$ *denote a rectangular matrix,* $\mathbf{\Sigma_H} \in \mathcal{C}^{n\times n}$ *denote a positive semi-definite matrix, and* $\mathbf{P}^\dagger\mathbf{\Sigma_H}\mathbf{P} \in \mathcal{C}^{m\times m}$ *be a diagonal matrix with diagonal elements in decreasing order (possibly with some zero diagonal elements). Then, there is a matrix of the form* $\overline{\mathbf{P}} = \mathbf{V_H}\ [\ \mathbf{\Lambda},\ \mathbf{0}\ ]$ *that satisfies:*

$$\overline{\mathbf{P}}^\dagger\mathbf{\Sigma_H}\overline{\mathbf{P}} = \alpha\mathbf{P}^\dagger\mathbf{\Sigma_H}\mathbf{P} \tag{39}$$

$$\text{tr}(\overline{\mathbf{P}}\ \overline{\mathbf{P}}^\dagger) = \text{tr}(\mathbf{P}\mathbf{P}^\dagger) \tag{40}$$

where $\alpha \geq 1$, $\mathbf{V_H}$ *is a unitary matrix with columns equal to the eigenvectors of matrix* $\mathbf{\Sigma_H}$ *corresponding to the* $\min(n,m)$ *largest eigenvalues in decreasing order and* $\mathbf{\Lambda}$ *is square diagonal matrix of size* $\min(n,m)$.

Proof: This is a modification of Lemma 3.16 in [47]. ■

Lemma 6. *For the complex gradient defined in* (35) *and general matrices* $\mathbf{A} \in \mathbb{C}^{m\times m}$, $\mathbf{B} \in \mathbb{C}^{n\times n}$ *and* $\mathbf{X} \in \mathbb{C}^{m\times n}$, *we have:*

$$\nabla_{\mathbf{X}}\text{tr}(\mathbf{A\ X\ B\ X}^\dagger) = \mathbf{A\ X\ B} \tag{41}$$

Proof: Using properties of differentials (32) and (33) from [48] we have:

$$\partial\text{tr}(\mathbf{A\ X\ B\ X}^\dagger) = \text{tr}[\mathbf{A}\ \partial(\mathbf{X})\ \mathbf{B\ X}^\dagger] + \text{tr}[\mathbf{A\ X\ B}\ \partial(\mathbf{X}^\dagger)]. \tag{42}$$

Together with the results for complex derivatives (219), (220), (221) and (222) from [48] we have:

$$\frac{\partial}{\partial\texttt{Re}(\mathbf{X})}\text{tr}(\mathbf{A\ X\ B\ X}^\dagger) = \mathbf{A}^\intercal\ \mathbf{X}^*\ \mathbf{B}^\intercal + \mathbf{A\ X\ B} \tag{43}$$

$$i\frac{\partial}{\partial\texttt{Im}(\mathbf{X})}\text{tr}(\mathbf{A\ X\ B\ X}^\dagger) = -\mathbf{A}^\intercal\ \mathbf{X}^*\ \mathbf{B}^\intercal + \mathbf{A\ X\ B} \tag{44}$$

and the result follows. ■

Remark 4. *This is the counterpart for complex-valued matrices to result (108) in [48] for real-valued matrices; note that the term* $\mathbf{A}^\intercal\ \mathbf{X}\ \mathbf{B}^\intercal$ *is absent in the complex case.*

Lemma 7. *For the complex gradient defined in* (35) *and general matrices* $\mathbf{A} \in \mathbb{C}^{m\times m}$, $\mathbf{B} \in \mathbb{C}^{n\times n}$, $\mathbf{C} \in \mathbb{C}^{n\times n}$ *and* $\mathbf{X} \in \mathbb{C}^{m\times n}$, *we have:*

$$\begin{aligned}&\nabla_{\mathbf{X}}\text{tr}\left[(\mathbf{A} + \mathbf{X\ B\ X}^\dagger)^{-1}\ (\mathbf{X\ C\ X}^\dagger)\right] \\ &= (\mathbf{A} + \mathbf{XBX}^\dagger)^{-1}\ \mathbf{X\ C}\left[\mathbf{I} - \mathbf{X}^\dagger\ (\mathbf{A} + \mathbf{XBX}^\dagger)^{-1}\ \mathbf{XB}\right]\end{aligned} \tag{45}$$

Proof: Using properties of differentials (32), (33) and (36) in [48] and the abbreviation $\mathbf{Y} = \mathbf{A} + \mathbf{X\ B\ X}^\dagger$, we have:

$$\begin{aligned}\partial\text{tr}\left[\mathbf{Y}^{-1}\ (\mathbf{X\ C\ X}^\dagger)\right] &= \text{tr}\left\{\mathbf{Y}^{-1}\ \partial(\mathbf{X\ C\ X}^\dagger)\right\} \\ &+ \text{tr}\left\{-\mathbf{Y}^{-1}\partial(\mathbf{A} + \mathbf{X\ B\ X}^\dagger)\mathbf{Y}^{-1}\ (\mathbf{X\ C\ X}^\dagger)\right\}\end{aligned}$$

Applying Lemma 6 the result follows. ■

Remark 5. *This is the counterpart for complex-valued matrices to result (116) in [48] for real-valued matrices; note that we do not require the assumption that* $\mathbf{B}$ *and* $\mathbf{C}$ *are Hermitian (symmetric) and there is no factor of* 2.

Appendix C
Proof of Theorem 1

Proof: We first provided an alternative derivation of this proof in [17]. The current proof is derived directly from the proof for an identical theorem for radar in [32]. We restate the mutual information between the input and output of the compressive sensing model in (1) for a multivariate Gaussian source as follows:

$$\mathcal{I}(\mathbf{x};\mathbf{y}) = \log\det\left(\mathbf{I}_m + \mathbf{M}^\dagger\mathbf{\Sigma}_{\mathbf{w}}^{-1}\mathbf{M}\ \mathbf{\Sigma_x}\right). \tag{11}$$

Note that for a unitary matrix $\mathbf{U}$, the kernel $\mathbf{P} = \mathbf{MU}$ has the same power as $\mathbf{M}$, i.e., $\text{tr}(\mathbf{PP}^\dagger) = \text{tr}(\mathbf{MM}^\dagger)$, but it may have different mutual information. In particular, a choice of $\mathbf{U}$ that maximizes the mutual information for a given $\mathbf{M}$ is $\mathbf{U} = \mathbf{U_x}$. This can be seen from Lemma 4 and Remark 4. From Lemma 5 we know that there exists a matrix $\overline{\mathbf{P}} = \mathbf{U_w}\ [\ \mathbf{\Lambda},\ \mathbf{0}\ ]$, which satisfies $\text{tr}(\overline{\mathbf{P}}\ \overline{\mathbf{P}}^\dagger) = \text{tr}(\mathbf{PP}^\dagger)$ and $\overline{\mathbf{P}}^\dagger\mathbf{\Sigma_H}\overline{\mathbf{P}} = \alpha\ \mathbf{P}^\dagger\mathbf{\Sigma}_{\mathbf{w}}^{-1}\mathbf{P}$ where $\alpha \geq 1$. Since the function $\det(\mathbf{I} + \alpha\mathbf{A})$ is monotonically increasing in α for a positive semi-definite matrix $\mathbf{A}$, the optimal kernel matrix must have the form of $\mathbf{M}^\star = \mathbf{U}_{\mathbf{M}}^\star\mathbf{\Lambda}_{\mathbf{M}}^\star\mathbf{V}_{\mathbf{M}}^{\star\dagger} = \mathbf{U_w}\ [\ \mathbf{\Lambda},\ \mathbf{0}\ ]\mathbf{U}_{\mathbf{x}}^\dagger$.

Finally, we determine the optimal singular values by optimizing the mutual information with respect to the eigenvalues rather than the singular values, since 1) they map one-to-one (up to a factor of $\exp j\theta$, which does not affect the mutual information) and 2) this new optimization problem is convex, so the Karush-Kuhn-Tucker (KKT) optimality conditions [49] define the unique global optimum. This is given by:

$$\lambda_{M_i}^\star = \begin{cases}0, & \frac{1}{\eta} - \frac{\lambda_{w_i}}{\lambda_{x_i}} \leq 0, \\ \frac{1}{\eta} - \frac{\lambda_{w_i}}{\lambda_{x_i}}, & \frac{1}{\eta} - \frac{\lambda_{w_i}}{\lambda_{x_i}} > 0\end{cases} \tag{46}$$

where η is such that the average unit norm row constraint is satisfied, i.e., $\frac{1}{\ell}\sum\lambda_{M_i}^\star = 1$, where the eigenvalues of $\mathbf{\Sigma_x}$ are arranged in descending order and the eigenvalues of $\mathbf{\Sigma_w}$ are arranged in ascending order , i.e., $\lambda_{x1} \geq \cdots \geq \lambda_{xm} \geq 0$ and $0 \leq \lambda_{w1} \leq \cdots \leq \lambda_{w_\ell}$. ■

Appendix D
Proof of Theorem 2

Proof: The proof draws on the work by Payaró and Palomar [36], which described the generalized mercury waterfilling aspect of the solution, but not the mode alignment aspect, and the work by Lamarca [24], which described the mode alignment aspect but did not focus on the generalized mercury waterfilling interpretation. The current proof highlights both the mode alignment and mercury waterfilling aspects of the solution.

The solution to the optimization problem in (10) satisfies the KKT optimality conditions:

$$\nabla_{\mathbf{M}} \left\{ -\mathcal{I}\left(\mathbf{x};\mathbf{y}\right) - \eta \cdot \left[\ell - \mathsf{tr}\left(\mathbf{M}\mathbf{M}^{\dagger}\right) \right] \right\} \Big|_{\mathbf{M}=\mathbf{M}^{\star}} = 0 \quad (47)$$

$$\eta \cdot \left[\ell - \mathsf{tr}\left(\mathbf{M}^{\star}\mathbf{M}^{\star\dagger}\right) \right] = 0 \quad (48)$$

with $\eta \geq 0$. Using the relationship between the gradient of the mutual information and the MMSE matrix in (7), the optimal kernel satisfies:

$$\eta \cdot \mathbf{M}^{\star}\mathbf{M}^{\star\dagger} = \mathbf{\Sigma}_{\mathbf{w}}^{-1} \ \left(\mathbf{M}^{\star} \ \mathbf{E}^{\star} \ \mathbf{M}^{\star\dagger}\right). \quad (49)$$

We note that (49) is diagonalized by $\mathbf{U}_{\mathbf{M}}^{\star}$, by definition, from which it can be seen that the matrices $\mathbf{\Sigma}_{\mathbf{w}}^{-1}$ and $\mathbf{M}^{\star}\mathbf{E}^{\star}\mathbf{M}^{\star\dagger}$ commute. From this observation, together with the fact that (49) is Hermitian and Lemma 2, we deduce that:

$$\mathbf{U}_{\mathbf{M}}^{\star} = \mathbf{U}_{\mathbf{w}}\mathbf{\Pi}_{\mathbf{U}}^{\star}\mathbf{\Lambda}_{\mathbf{U}} \quad (50)$$

where $\mathbf{\Lambda}_{\mathbf{U}}$ is a diagonal matrix with unit modulus diagonal elements and $\mathbf{\Pi}_{\mathbf{U}}^{\star}$ is a permutation matrix. Furthermore, $\mathbf{U}_{\mathbf{M}}^{\star\dagger}\mathbf{M}^{\star}\mathbf{E}^{\star}\mathbf{M}^{\star\dagger}\mathbf{U}_{\mathbf{M}}^{\star}$ is a diagonal matrix, from which we can infer:

$$\mathbf{V}_{\mathbf{M}}^{\star} = \mathbf{U}_{\mathbf{E}}^{\star}\mathbf{\Pi}_{\mathbf{V}}^{\star}\mathbf{\Lambda}_{\mathbf{V}} \quad (51)$$

where $\mathbf{\Lambda}_{\mathbf{V}}$ is a diagonal matrix with unit modulus diagonal elements and $\mathbf{\Pi}_{\mathbf{V}}^{\star}$ is a permutation matrix. Both mutual information and the MMSE matrix are independent of $\mathbf{\Lambda}_{\mathbf{U}}$ and $\mathbf{\Lambda}_{\mathbf{V}}$, allowing us to write without loss of generality the optimal unitary matrices as follows:

$$\mathbf{U}_{\mathbf{M}}^{\star} = \mathbf{U}_{\mathbf{w}} \quad (52)$$

$$\mathbf{V}_{\mathbf{M}}^{\star} = \mathbf{U}_{\mathbf{E}}^{\star} \ \mathbf{\Pi}^{\star} \quad (53)$$

where $\mathbf{\Pi}^{\star}$ is some optimal permutation matrix.

By setting $\mathbf{U}_{\mathbf{M}}^{\star} = \mathbf{U}_{\mathbf{w}}$ we can now obtain an equivalent[8] channel model:

$$\mathbf{y}' = \mathbf{\Lambda}_{\mathbf{w}}^{-1/2} \ \mathbf{\Lambda}_{\mathbf{M}} \ \mathbf{V}_{\mathbf{M}}^{\dagger} \ \mathbf{x} + \mathbf{n} \quad (54)$$

where $\mathbf{y}' = \mathbf{\Lambda}_{\mathbf{w}}^{-1/2} \ \mathbf{U}_{\mathbf{w}}^{\dagger} \ \mathbf{y}$ and $\mathbf{n} = \mathbf{\Lambda}_{\mathbf{w}}^{-1/2} \ \mathbf{U}_{\mathbf{w}}^{\dagger} \ \mathbf{w}$ is zero-mean circularly symmetric complex Gaussian noise with identity covariance, $\mathbf{n} \sim \mathcal{CN}(\mathbf{n};\mathbf{0},\mathbf{I})$.

It was shown in [26] that for a fixed value of $\mathbf{V}_{\mathbf{M}}$ the mutual information $\mathcal{I}(\mathbf{x};\mathbf{y}')$ is concave with respect to the squared singular values of $\mathbf{\Lambda}_{\mathbf{M}}$, i.e., the following optimization problem has a unique global optimum given by the KKT conditions, where $\mathbf{\Lambda}_{\mathbf{Q}} = \mathbf{\Lambda}_{\mathbf{M}} \ \mathbf{\Lambda}_{\mathbf{M}}^{\dagger}$:

$$\begin{aligned} \underset{\lambda_{M_1},\lambda_{M_2},\ldots,\lambda_{M_\ell}}{\text{maximize}} \quad & \mathcal{I}\left(\mathbf{x};\mathbf{y}'\right) \\ \text{subject to} \quad & \sum_{i=1}^{\ell} \lambda_{M_i} \leq \ell \\ & \lambda_{M_i} \geq 0 \end{aligned} \quad (55)$$

The Lagrangian for this optimization problem is:

$$\mathcal{L}(\mathbf{\Lambda}_{\mathbf{Q}}) = \mathcal{I}(\mathbf{x};\mathbf{y}') + \eta\left(\ell - \sum_{i=1}^{\ell}\lambda_{M_i}\right) + \sum_{i=1}^{\ell}\eta_i\lambda_{M_i} \quad (56)$$

and the Karush-Kuhn-Tucker conditions state that:

$$\frac{\partial}{\partial \mathbf{\Lambda}_{\mathbf{Q}}}\mathcal{L}(\mathbf{\Lambda}_{\mathbf{Q}})\Big|_{\mathbf{\Lambda}_{\mathbf{Q}}=\mathbf{\Lambda}_{\mathbf{Q}}^{\star}} = 0 \quad (57)$$

$$\eta \cdot \left(\ell - \sum_{i=1}^{\ell}\lambda_{M_i}^{\star}\right) = 0 \quad (58)$$

$$\eta_i \cdot \lambda_{M_i}^{\star} = 0, \qquad i = 1,\ldots,\ell \quad (59)$$

By using the result from [26] that states that:

$$\frac{\partial}{\partial \mathbf{\Lambda}_{\mathbf{Q}}}\mathcal{I}\left(\mathbf{x};\mathbf{y}'\right) = \mathsf{Diag}\left(\mathbf{E}_{\mathbf{x}'} \ \mathbf{\Lambda}_{\mathbf{w}}^{-1} \ \right) \quad (60)$$

where $\mathbf{E}' = \mathbf{V}_{\mathbf{M}}^{\dagger} \ \mathbf{E} \ \mathbf{V}_{\mathbf{M}}$ is the MMSE matrix associated with the estimation of $\mathbf{x}' = \mathbf{V}_{\mathbf{M}}^{\dagger} \ \mathbf{x}$ from $\mathbf{y}'$, it is possible to rewrite (57) as follows:

$$\eta\lambda_{w_i} - \mathsf{mmse}_i\left(\mathbf{V}_{\mathbf{M}},\mathbf{\Lambda}_{\mathbf{Q}}^{\star}\right) = \eta_i\lambda_{w_i}, \qquad i = 1,\ldots,\ell. \quad (61)$$

where $\mathsf{mmse}_i\left(\mathbf{V}_{\mathbf{M}},\mathbf{\Lambda}_{\mathbf{Q}}\right)$ denotes the i-th diagonal entry of $\mathbf{E}_{\mathbf{x}'}$ for that particular $\mathbf{\Lambda}_{\mathbf{Q}}$.

From the KKT conditions, we know that if $\lambda_{M_i}^{\star} > 0$ then $\eta_i = 0$ and that $\eta > 0$. For a given value of η, the value of $\lambda_{M_i}^{\star}$ can be calculated from the relationship $\eta \ \lambda_{w_i} = \mathsf{mmse}_i\left(\mathbf{V}_{\mathbf{M}},\mathbf{\Lambda}_{\mathbf{Q}}^{\star}\right)$ for fixed $\lambda_{M_j},\forall j \neq i$. The function mmse_i is non-negative and monotonically decreasing in $\lambda_{M_i} \in [0,\infty]$ for fixed $\lambda_{M_j},\forall j \neq i$, and its maximum value is given when $\lambda_{M_i} = 0$. Therefore if $\eta \ \lambda_{w_i} > \mathsf{mmse}_i\left(\mathbf{V}_{\mathbf{M}},\mathbf{\Lambda}_{\mathbf{Q}}^{\star}|_{\lambda_{M_i}^{\star}=0}\right)$ where $\mathbf{\Lambda}_{\mathbf{Q}}^{\star}|_{\lambda_{M_i}^{\star}=0} = \mathsf{Diag}\left(\lambda_{M_1}^{\star},\ldots,\lambda_{M_{i-1}}^{\star},0,\lambda_{M_{i+1}}^{\star},\ldots,\lambda_{M_\ell}^{\star}\right)$, then $\lambda_{M_i}^{\star} = 0$ and $\eta_i \neq 0$.

This result is true for all values of $\mathbf{V}_{\mathbf{M}}$, therefore it is also true when $\mathbf{V}_{\mathbf{M}} = \mathbf{U}_{\mathbf{E}}^{\star} \ \mathbf{\Pi}^{\star}$ and the result follows. ■

Appendix E
Proof of Theorem 3

Proof: Note that:

$$p(\mathbf{y}) = \mathcal{CN}\left(\mathbf{y};\mathbf{M} \ \mathbf{x},\mathbf{\Sigma}_{\mathbf{w}} + \mathbf{M} \ \mathbf{\Sigma}_{\mathbf{x}} \ \mathbf{M}^{\dagger}\right) \quad (62)$$

and so:

$$p^{\alpha}(\mathbf{y}) = \frac{\mathcal{CN}\left(\mathbf{y};\mathbf{M} \ \mathbf{x},\frac{1}{\alpha}(\mathbf{\Sigma}_{\mathbf{w}} + \mathbf{M} \ \mathbf{\Sigma}_{\mathbf{x}} \ \mathbf{M}^{\dagger})\right)}{\alpha^{k}\left(2\pi\right)^{k(\alpha-1)}\det\left(\mathbf{\Sigma}_{\mathbf{w}} + \mathbf{M} \ \mathbf{\Sigma}_{\mathbf{x}} \ \mathbf{M}^{\dagger}\right)^{\alpha-1}} \quad (63)$$

[8] The equivalence is in the sense that the mutual information between the input and the output of both models is equal, i.e., $\mathcal{I}\left(\mathbf{x};\mathbf{y}\right) = \mathcal{I}\left(\mathbf{x};\mathbf{y}'\right)$.

By substituting this into the expression for Rényi entropy it follows that:

$$\mathrm{h}_{\alpha}(\mathbf{y}) = \log\left[(2\pi)^{k}\det(\mathbf{\Sigma}_{\mathbf{w}} + \mathbf{M}\ \mathbf{\Sigma}_{\mathbf{x}}\ \mathbf{M}^{\dagger})\right] + \frac{k\log\alpha}{(\alpha-1)}. \tag{64}$$

The result now follows using the definition of Shannon entropy for Gaussian sources. ■

Appendix F
Proofs for Gradient of Rényi entropy

Proof: Let us first show that we can express the following relevant gradient analytically:

$$\begin{aligned}
&\nabla_{\mathbf{M}}\log\mathcal{CN}\left(\mathbf{0};\boldsymbol{\mu}_{i,j},\mathbf{\Sigma}_{i,j}\right) = \\
&-\nabla_{\mathbf{M}}\left(k\log 2\pi\right) - \nabla_{\mathbf{M}}\left\{\log\det\mathbf{\Sigma}_{i,j}\right\} && (65)\\
&-\nabla_{\mathbf{M}}\left\{\boldsymbol{\mu}_{i,j}^{\intercal}\mathbf{\Sigma}_{i,j}^{-1}\boldsymbol{\mu}_{i,j}\right\}, && (66)
\end{aligned}$$

where $\boldsymbol{\mu}_{i,j} = \mathbf{M}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)$ and $\mathbf{\Sigma}_{i,j} = \mathbf{M}\left(\mathbf{\Sigma}_i + \mathbf{\Sigma}_j\right)\mathbf{M}^{\dagger} + 2\mathbf{\Sigma}_{\mathbf{w}}$.

The first term is zero, the second term is the mutual information for a complex Gaussian distribution and can be evaluated using (7), relating the mutual information and MMSE matrix:

$$\nabla_{\mathbf{M}}\log\det\mathbf{\Sigma}_{i,j} = (2\mathbf{\Sigma}_{\mathbf{w}})^{\mathbf{-1}}\mathbf{M}\ \mathbf{E}_{\mathbf{i,j}} \tag{67}$$

where the $\mathbf{E}_{i,j} = \left[\left(\mathbf{\Sigma}_i + \mathbf{\Sigma}_j\right)^{-1} + \mathbf{M}^{\dagger}(2\mathbf{\Sigma}_{\mathbf{w}})^{\mathbf{-1}}\mathbf{M}\right]^{-1}$ is the MMSE matrix if the input signal x was Gaussian distributed with covariance $\left(\mathbf{\Sigma}_i + \mathbf{\Sigma}_j\right)$ and distorted by Gaussian noise with covariance $2\mathbf{\Sigma}_{\mathbf{w}}$. It can also be expressed:

$$\nabla_{\mathbf{M}}\log\det\mathbf{\Sigma}_{i,j} = \mathbf{\Sigma}_{i,j}^{-1}\ \mathbf{M}\left(\mathbf{\Sigma}_i + \mathbf{\Sigma}_j\right) \tag{68}$$

where we can use Woodbury's Inversion Lemma to convert between the two. The third and final term, using Lemma 7 and chain rule (94) in [28], can be expressed:

$$\begin{aligned}
&\nabla_{\mathbf{M}}\left\{\boldsymbol{\mu}_{i,j}^{\intercal}\mathbf{\Sigma}_{i,j}^{-1}\boldsymbol{\mu}_{i,j}\right\} = \ \mathbf{\Sigma}_{i,j}^{-1}\ \mathbf{M}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^{\dagger}\\
&\times\left\{\mathbf{I} - \mathbf{M}^{\dagger}\mathbf{\Sigma}_{i,j}^{-1}\mathbf{M}\left(\mathbf{\Sigma}_i + \mathbf{\Sigma}_j\right)\right\}. && (69)
\end{aligned}$$

■