# Information-Theoretic Compressive Measurement Design

Liming Wang, Minhua Chen, Miguel Rodrigues, David Wilcox, Robert Calderbank, and Lawrence Carin

**Abstract**—An information-theoretic projection design framework is proposed, of interest for feature design and compressive measurements. Both Gaussian and Poisson measurement models are considered. The gradient of a proposed information-theoretic metric (ITM) is derived, and a gradient-descent algorithm is applied in design; connections are made to the information bottleneck. The fundamental solution structure of such design is revealed in the case of a Gaussian measurement model and arbitrary input statistics. This new theoretical result reveals how ITM parameter settings impact the number of needed projection measurements, with this verified experimentally. The ITM achieves promising results on real data, for both signal recovery and classification.

**Index Terms**—Information-theoretic metric, information bottleneck, projection design, gradient of mutual information, compressive sensing

## 1 Introduction

Dimensionality reduction plays a pivotal role in many machine-learning applications, including compressive measurements [1] and feature design [2], [3], [4]. Although there is interest in nonlinear representations [5], linear compressive measurements and feature design are desirable because they are amenable to theoretical analysis and guarantees [6], and such representations can also be readily implemented experimentally [7]. Linear dimensionality reduction is achieved by multiplying a signal of interest with a measurement matrix, and the number of rows of this matrix defines the number of measurements/features; that number is ideally small relative to the dimension of the original data vector. One typically assumes additive noise in the measurement, which is often modeled as being Gaussian [8], [9]. However, there are low-photon-count applications for which a Gaussian noise model is inappropriate, and a Poisson measurement model is desired. There are far fewer theoretical results for the Poisson case.

Among the various criteria for measuring the information in a compressive measurement or in feature design, mutual information is widely used [10], [11], due to its ubiquitous presence as an information loss measure, as well as its close relationship to Bayesian classification error [12]. Mutual-information-based measurement design for the Gaussian model has been considered in [8] for signal recovery, and in [4] for classification (feature design). If one designs compressive measurements based only on a classification criterion, the features may be good for *an algorithm* to classify, but a human may also wish to look at the underlying data (e.g., a medical scientist may wish to confirm an algorithm-defined diagnosis), and poor signal recovery may undermine that objective. It is therefore of interest to jointly balance the goals of feature design (for classification) and signal-recovery quality.

In this paper, we propose an information-theoretic metric (ITM) by balancing two performance metrics linked to the data. The proposed ITM is able to accommodate non-Gaussian inputs and Poisson measurement noise. Further, we also consider two types of constraints for the ITM, an energy and orthonormality constraint (on the characteristics of the projection vectors), which are widely utilized in many applications.

Under proper settings of the model, the proposed ITM is like the information bottleneck (IB) [13], which has been widely used in various applications, including document clustering [14], gene-expression analysis [15], video search [16], feature design [17] and speech recognition [18], among many others. The case for which the input signal is Gaussian and there is Gaussian measurement noise has been investigated in [19], where an analytic solution of the IB problem has been revealed.

In situations for which the input signal and/or the measurement is *not* Gaussian, the mutual information terms involved in the ITM generally do not possess analytic forms under almost all common input distributions. A goal of this paper is to develop numerical algorithms as well as theoretical results for projection-matrix design under the ITM, for general Gaussian and Poisson measurement models, for non-Gaussian input signals. Despite the lack of analytic mutual information expressions, the gradients of mutual information for both Gaussian and Poisson measurement models possess relatively simple forms, which have


- *L. Wang, R. Calderbank and L. Carin are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708. E-mail: {liming.w, robert.calderbank, lcarin}@duke.edu.*
- *M. Chen is with the Department of Computer Science and Statistics, University of Chicago, Chicago, IL 60637. E-mail: minhua@cs.uchicago.edu.*
- *D. Wilcox is with the Department of Chemistry, Purdue University, West Lafayette, IN 47907. E-mail: wilcoxd@purdue.edu.*
- *M. Rodrigues is with the Department of Electronic and Electrical Engineering, University College London, London WC1E 7JE, United Kingdom. E-mail: m.rodrigues@ee.ucl.ac.uk.*


*Manuscript received 6 May 2014; revised 9 Apr. 2016; accepted 9 May 2016. Date of publication 12 May 2016; date of current version 12 May 2017. Recommended for acceptance by M. A. Carreira-Perpinan. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TPAMI.2016.2568189*

connections to the Minimum Mean Square Error (MMSE) estimator [8], [20], [21], [22], [23]. Hence, gradient-descent methods may be adopted to infer a desired measurement matrix. In the context of compressive sensing, this paper is believed to be the first for which a balance between signal recovery and classification is considered.

The ITM setup has a parameter $\beta$ that may be tuned, which weighs between the classification and signal-recovery objectives. As detailed below, for $\beta < 0$ we jointly consider the goals of classification and recovery; with $\beta \to -\infty$ we optimize only for signal recovery, and when $\beta = 0$ we optimize only for classification. For $\beta > 0$ the ITM considers classification with a penalty ("bottleneck") imposed on signal recovery. For $\beta > 0$, our new theory demonstrates that larger $\beta$ encourages fewer number of linear measurements/features used in the classifier. This result for $\beta > 0$ under the energy constraint has an interesting waterfilling-like interpretation, generalizing the mode alignment-scheme in [8], [24] to the IB. Further, the derived waterfilling-like interpretation for ITM under the energy constraint is valid for arbitrary source model and extends the result in [19].

The ITM under the orthonormality constraint is related to calculation of volumes on the Stiefel manifold, where the derivation of an analytical solution is known to be challenging [25]. We shed light on its analytical solution by postulating a Gaussian input and Gaussian noise with diagonal variance as in [19], and a fixed-point-type solution is derived, for which analytic solutions are realized for $\beta \to -\infty$ and $\beta = 0$. While aspects of this theory is limited to the case of a Gaussian source, the experiments (numerical methods) are not.

We demonstrate how the theoretical results may be used in practice, for projection design or supervised dimensionality reduction. Specifically, we provide numerical results on several datasets, and compare the performance of the proposed ITM to Linear Discriminant Analysis (LDA) [26], Information Discriminant Analysis (IDA) [12] and a Rényi-entropy method [2], [3]. For a Gaussian measurement model, limiting cases of the ITM correspond to special cases in [8] and [4], to which we also compare. For the Poisson measurement model, we apply the ITM on real measured data from a compressive photon-counting hyperspectral camera (the measurement system in [27]), and demonstrate that designed projection matrices can yield improved performance on preserving both signal recovery and classification information, relative to random measurement design [28].

The remainder of the paper is organized as follows. In Section 2, we provide a formal definition of the proposed ITM and related optimization problems of interest. We derive explicit forms of the gradients of the ITM in Section 3, and summarize a gradient-based numerical algorithm for the ITM. In Section 4 we present two theorems, on the solution structure for the ITM under energy and orthonormality constraints on the measurement matrix, and we elaborate on the interpretations of these theoretical results. We present experimental results for the ITM in Section 5, for both Gaussian and Poisson measurement models; for the Poisson case, the methods are applied to real data from a measurement system. Finally, we conclude the paper in Section 6.

## 2 Information-Theoretic Framework for Designed Measurements

### 2.1 Problem Statements

The setup we consider is characterized by the Markov sequence $X \to Y \to \tilde{Y} \to \hat{X}$. Two types of signal models are considered. For the first, *discrete* $X \in \{1, \ldots, L\}$ denotes the class label, and $Y \in \mathbb{R}^n$ are data conditioned on $X$. Alternatively, for the second signal model, the discrete $X$ in the Markov sequence is replaced by a *continuous* random variable $X' \in \mathbb{R}^p$, and $(X', Y)$ are assumed to be jointly Gaussian. $\tilde{Y} \in \mathbb{R}^m$ is the compressive form of $Y$ with $m \ll n$, and $\hat{X}$ is the estimated $X$ (or $\hat{X}'$ is the estimated $X'$), based on $\tilde{Y}$.

For the first model, $p_X(X = i) = \pi_i$, with $\pi_i > 0$ and $\sum_{i=1}^{L} \pi_i = 1$. The random variable $Y$ is modeled as drawn from a mixture model: $P_Y = \sum_{i=1}^{L} P_X(X = i) P_{Y|X}(Y|X = i) = \sum_{i=1}^{L} \pi_i P_{Y|X}(Y|X = i)$. We consider general conditional distributions $P_{Y|X}$.

The mapping $Y \to \tilde{Y}$ is stochastic, and is here effected first in terms of the *deterministic* linear mapping $\Phi Y$, where $\Phi \in \mathbb{R}^{m \times n}$ is a measurement matrix. For both the Gaussian and Poisson measurement models that are the focus of this paper, the $m$-dimensional signal $\Phi Y$ defines the mean of the mapping $Y \to \tilde{Y}$. Specifically, for the Gaussian measurement channel

$$P_{\tilde{Y}|Y} = \mathcal{N}(\tilde{Y}; \Phi Y, \Lambda^{-1}), \tag{1}$$

where $\Lambda \in \mathbb{R}^{m \times m}$ is the precision of the additive noise in the Gaussian measurement.

For the Poisson measurement model

$$P_{\tilde{Y}|Y} = \mathrm{Pois}(\tilde{Y}; \Phi Y + \lambda) = \prod_{i=1}^{m} \mathrm{Pois}(\tilde{Y}_i; (\Phi Y)_i + \lambda_i), \tag{2}$$

where $\Phi \in \mathbb{R}_+^{m \times n}$, $\lambda \in \mathbb{R}_+^m$ is the "dark current," and $Y_i$ represents component $i$ of vector $Y$ (this subscript notation identifies vector components throughout the paper). For the Poisson case we also require that the components of $Y$ are non-negative. One may view $Y \in \mathbb{Z}_+^n$ as the $n$-dimensional Poisson rate of the original signal, and $\Phi Y$ is the compressive form of the Poisson rate. In (2), $\mathrm{Pois}(\tilde{Y}; \Phi Y + \lambda)$ denotes a vector Poisson distribution, while $\mathrm{Pois}(\tilde{Y}_i; (\Phi Y)_i + \lambda_i)$ is the common scalar Poisson distribution; the form of the distribution is assumed understood from the context in which it is used, to not complicate notation.

For the second model discussed above, $(X', Y)$ are drawn jointly Gaussian, and therefore $Y$ is not restricted to be non-negative. Hence, the second model is restricted to the case in (1), the Gaussian measurement model. If interested in the Poisson measurement model, $Y$ may be treated as the log of the Poisson rate.

The proposed ITM is characterized by a linear combination of mutual information $I(X; \tilde{Y})$ and $I(Y, \tilde{Y})$; Shannon entropy with the natural logarithm is considered below. The ITM setup for defining $\Phi$ is

$$\max_{\Phi} \mathcal{ITM}(\Phi, \beta) := \max_{\Phi} \big\{ I(X; \tilde{Y}) - \beta I(\tilde{Y}, Y) \big\}, \tag{3}$$

where $\beta \in \mathbb{R}$ controls the relative importance of the two mutual-information metrics.

We consider two different types of constraints on $\Phi$. The first is the energy constraint:

$$\frac{1}{m}\mathrm{tr}(\Phi\Phi^T) \leq 1, \tag{4}$$

where the trace $\mathrm{tr}(\cdot)$ manifests an energy regularization on the measurement. The second constraint imposes that the projections are orthonormal:

$$\Phi\Phi^T = I_m, \tag{5}$$

where $I_m$ is the $m \times m$ identity matrix.

In compressive sensing, one often requires that $\Phi$ obey unit-norm row constrains or, instead, orthonormality constraints [12]. For communications applications, one may desire the energy constraint, requiring the rows have unit-norm on average [29].

We consider the first signal model (discrete $X$) unless otherwise specified, for both the Gaussian and Poisson measurement models. It is assumed that $P_X$ and $P_{Y|X}$ are known *a priori*. However, we do not assume any specific form of $P_{Y|X}$, thus a general mixture model is considered for $Y$. On the other hand, for the second signal model, for which $(X', Y)$ are jointly Gaussian, the ITM for the Gaussian measurement model possesses an analytical expression, thereby enabling a direct optimization and analysis [8], [13].

### 2.2 Mutual Information Metrics

The mapping $Y \rightarrow \tilde{Y}$ constitutes the heart of the compressive measurement or linear feature design, and we wish to design $\Phi$ such that it preserves desired information. The mutual information is employed as in (3), from two perspectives, which we motivate here.

In [8] the authors considered compressive sensing, and the goal was to recover $Y$ from $\tilde{Y}$. The measurement model $P_{\tilde{Y}|Y}$ was assumed to be Gaussian as in (1), and $P_Y$ was arbitrary. In the experiments in [8], a mixture model was used for $Y$: $P_Y = \sum_{i=1}^{L} P_{X=i} P_{Y|X=i}$; in [8] the focus was not on recovering $X$, but rather on estimating $Y$. In that case the mutual information of interest is $I(Y;\tilde{Y})$, and the goal is to design $\Phi$ such that $I(Y;\tilde{Y})$ is maximized. This design framework may be justified by noting that it has been shown recently that [30]

$$\mathrm{MMSE} \geq \frac{1}{2\pi e}\exp\{2[h(Y) - I(\tilde{Y};Y)]\}, \tag{6}$$

where $h(Y)$ is the differential entropy of $Y$ and $\mathrm{MMSE} = \mathbb{E}\{\mathrm{tr}[(Y - \mathbb{E}(Y|\tilde{Y}))(Y - \mathbb{E}(Y|\tilde{Y}))^T]\}$ is the minimum mean-square error, so that by maximizing mutual information one may hope to achieve a lower reconstruction error.

For the classification problem considered in [4], the objective was to recover the class label $X$ from $\tilde{Y}$, and therefore the goal there was to design $\Phi$ to maximize $I(X;\tilde{Y})$. This is justified by recalling the Bayesian classification error $P_e = \int P_{\tilde{Y}}(\tilde{y})[1 - \max_X P_{X|\tilde{Y}}(x|\tilde{y})]d\tilde{y}$, and noting that it has been shown in [31] that

$$P_e \leq \frac{1}{2}H(X|\tilde{Y}), \tag{7}$$

where $H(X|\tilde{Y}) = H(X) - I(X;\tilde{Y})$, and $H(\cdot)$ denotes the entropy of a discrete random variable. Since $H(X)$ is independent of $\Phi$, minimizing the upper bound to $P_e$ is equivalent to maximizing $I(X;\tilde{Y})$.

There are practical settings for which *both* signal recovery and classification are of interest simultaneously, which motivates the proposed ITM in 3 when $\beta \leq 0$. Note that $\beta = 0$ corresponds to the case that only the classification task is of interest and $\beta \rightarrow -\infty$ optimizes only for the signal recovery.

### 2.3 Connections to the Information Bottleneck

When $\beta > 0$, the proposed ITM is like the IB and the goal is to recover $X$ from $\tilde{Y}$, but one wishes to constrain the number of measurements/features $m$, this defining the "bottleneck". In the language of the IB [13], $X$ is the "relevant" information we principally wish to recover; $Y \in \mathbb{R}^n$ is the input signal that depends on $X$, $\tilde{Y} \in \mathbb{R}^m$ is the compressive measurement or representation of $Y$, and $\hat{X}$ is the estimate of $X$, based on $\tilde{Y}$. We note, however, that the IB does not typically impose constraints on $\Phi$, like we do in (4) and (5).

Due to different settings in the model, our ITM setup in (3) differs from the one proposed in [13]. Most importantly, in [13] $\beta$ was assumed to be nonnegative, which implies that the goal is to design $\Phi$ to maximize $I(X;\tilde{Y})$, while simultaneously minimizing $I(\tilde{Y},Y)$; since the maximization of $I(X;\tilde{Y})$ and the minimization of $I(\tilde{Y},Y)$ does not generally occur simultaneously for some $\Phi$, the choice of $\beta$ defines the relative importance of these two metrics on the design of $\Phi$. In this manner the model effectively seeks a low-complexity model, which is effective at retaining the "relevant" information $X$, while reducing complexity through simplified $\tilde{Y}$, that may not be effective for recovering $Y$. Our motivation is expanded beyond cases typically associated with the IB model, as we will also allow $\beta < 0$, permitting consideration of the case of *joint* interest in signal recovery and classification.

In general, one cannot perform the optimization in (3) analytically, and therefore numerical procedures are required. Specifically, one must take the gradient of $\mathcal{ITM}(\Phi,\beta)$ with respect to $\Phi$, for specified choices of $\beta$. While $\mathcal{ITM}(\Phi,\beta)$ itself may be difficult to compute, expanding on recent analysis, one may express the gradient of $\mathcal{ITM}(\Phi,\beta)$ in closed form, amenable to optimization algorithms (e.g., gradient descent).

## 3 Gradients of the ITM Model

### 3.1 Explicit Forms for the Gradients

We introduce two theorems on gradients of the ITM, for both Gaussian and Poisson measurement models. We always assume the regularity conditions, specifically, that the order of integration and differentiation can be interchanged freely, e.g. the order of the differential operators $\frac{\partial}{\partial \Phi_{ij}}$ and the expectation operator $\mathbb{E}(\cdot)$ may be interchanged. This assumption is mild and almost always valid in practice [32].

The following two theorems are straightforward to establish given results from [4], [20], [21].

**Theorem 1.** *For the Gaussian measurement model in (1) under the first signal model, for arbitrary $P_{Y|X}$ such that the mixture distribution $P_Y$ is consistent with the assumed regularity conditions, the gradient of the ITM $\nabla_\Phi \mathcal{ITM}(\Phi,\beta)$ with respect to the projection matrix $\Phi$ is given by*

$$\nabla_\Phi \mathcal{ITM}(\Phi,\beta) = \Lambda\Phi\tilde{E} - \beta\Lambda\Phi E, \tag{8}$$

*where $E = \mathbb{E}\left[(Y-\mathbb{E}[Y|\tilde{Y}])(Y-\mathbb{E}[Y|\tilde{Y}])^T\right]$ is the* MMSE *matrix, and*

$$\tilde{E} = \mathbb{E}\left[(\mathbb{E}[Y|\tilde{Y},X]-\mathbb{E}[Y|\tilde{Y}])(\mathbb{E}[Y|\tilde{Y},X]-\mathbb{E}[Y|\tilde{Y}])^T\right].$$

**Theorem 2.** *For the Poisson measurement model in (2) under the first signal model, for arbitrary $P_{Y|X}$ such that the mixture distribution $P_Y$ is consistent with the assumed regularity conditions, the gradient of the ITM $\nabla_\Phi \mathcal{ITM}(\Phi,\beta)$ with respect to the projection matrix $\Phi$ is given by*

$$[\nabla_\Phi \mathcal{ITM}(\Phi,\beta)_{ij}] = \mathbb{E}\left[\mathbb{E}[Y_j|\tilde{Y},X]\log\frac{\mathbb{E}[(\Phi Y)_i|\tilde{Y},X]}{\mathbb{E}[(\Phi Y)_i|\tilde{Y}]}\right] - \beta\{\mathbb{E}\left[Y_j\log((\Phi Y)_i)\right] - \mathbb{E}\left[\mathbb{E}[Y_j|\tilde{Y}]\log\mathbb{E}[(\Phi Y)_i|\tilde{Y}]\right]\}. \tag{9}$$

*where $(\cdot)_{ij}$ denotes the $ij$th entry of the matrix and $(\cdot)_i$ denotes the $i$th entry of the vector.*

All theorem proofs are provided in the Appendix. Note that Theorems 1 and 2 are valid for *arbitrary* mixture model for $P_Y$ provided it satisfies the regularity conditions.

### 3.2 Gradient-Based Numerical Design

A numerical solution to the optimization in (3) can be realized via a gradient-descent method. The matrices $E$ and $\tilde{E}$ involved in Theorems 1 and 2 can be readily calculated by Monte Carlo integration; we elaborate on this calculation when presenting experimental results. The algorithm is summarized as follows:

1) Initialize $\Phi$ and either set or learn all the input distributions from the training datasets.
2) Use Monte Carlo integration to calculate the gradient. For Gaussian measurement model, the posteriors $P_{Y|\tilde{Y}}$ and $P_{Y|\tilde{Y},X}$ possess analytical expressions when signal $Y$ is assumed to be a Gaussian mixture model (GMM) [4]. Likewise, these posteriors may be readily approximated when the signal $Y$ is assumed to be a log-GMM, and the details are presented in Section 5.4. Therefore, the calculation of the gradients in Theorems 1 and 2 reduces to calculating the expectation $\mathbb{E}[f(\tilde{Y},X)]$ of some function $f$. It can be approximated via the Monte Carlo integration $\mathbb{E}[f(\tilde{Y},X)] \approx \frac{1}{S}\sum_{i=1}^{S} f((\tilde{y},x)_i)$, and $\{(\tilde{y},x)_i\}$, $i=1,\ldots,S$ are samples drawn from the distribution $P_{\tilde{Y},X}$, which can be obtained via the well-known Metropolis-Hastings algorithm [33].
3) Update the projection matrix as $\Phi^{new} = \mathrm{proj}(\Phi^{old} + \delta\nabla_\Phi \mathcal{ITM}(\Phi,\beta))$, where $\delta$ is the step size and $\mathrm{proj}(\cdot)$ projects the matrix to the feasible set either defined by the energy constraint in (4) or the orthonormality constraint in (5), i.e., re-normalize the matrix to satisfy the energy constraint or using Gram-Schmidt method to ortho-normalize the matrix to satisfy the orthonormality constraint.
4) Repeat previous step until convergence.

The above procedure (particularly the Monte Carlo integration) is attractive to implement for the specific forms of $P_{Y|X}$ assumed when presenting experimental results. In general, the ITM is neither a convex nor a concave function of $\Phi$, and therefore we are not guaranteed a global-optimal solution. In all experiments we have considered the solution converged to a useful/effective solution from a random start. When presenting experimental results, we also discuss how the positivity constraint on $\Phi$ is imposed for the Poisson measurement model.

## 4 Theoretical Analysis

In the previous section we proposed a numerical algorithm for design of the measurement matrix $\Phi$ based on the gradient-descent method, for solving the ITM construction in (3), with either constraint (4) or (5). We now investigate the theoretical characteristics of the solution structure. In [19], an analytic solution to the unconstrained IB model has been investigated, under the assumptions that $X'$ is a continuous random variable and $(X',Y)$ are jointly Gaussian (our *second* signal model), and a Gaussian measurement model has been assumed as well; by contrast, for our first signal model, of principal concern, $X$ is discrete. In this section, we first present a generalized water-filling-like [34], [35] solution to the Gaussian measurement model with the energy constraint as in (4). Specifically, as in [19] we assume a Gaussian measurement model, but we consider general statistics for $Y$, rather than assuming $P_Y$ is Gaussian (because $X$ is also not Gaussian, but is discrete).

Assuming the Gaussian measurement model in (1), we first define symbols that will be used in the statement of the theorem. Consider $E$ and $\tilde{E}$ as defined in Theorem 1. The covariance matrices of $X$, $X'$, $Y$, $Y|X$ and $Y|X'$ are denoted as $\Sigma_X$, $\Sigma_X'$, $\Sigma_Y$, $\Sigma_{Y|X}$ and $\Sigma_{Y|X'}$, respectively. Define $E'_\beta = \tilde{E} - \beta E$. Consider the singular value decomposition for $\Phi$, and eigenvalue decomposition for $E'_\beta$ and $\Sigma = \Lambda^{-1}$: $\Phi = U_\Phi D_\Phi V_\Phi^T$, $E'_\beta = U_{E'_\beta} D_{E'_\beta} U_{E'_\beta}^T$ and $\Sigma = U_\Sigma D_\Sigma U_\Sigma^T$, respectively, where $U_\Phi$, $V_\Phi$, $U_{E'_\beta}$ and $U_\Sigma$ are $m\times m$, $n\times n$, $n\times n$ and $m\times m$ orthonormal matrices, respectively. The eigenvalues in $D_\Phi$ and $D_{E'_\beta}$ are in descending order. The eigenvalues in $D_\Sigma$ are in ascending order, and we express $D_\Sigma = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_m^2)$.

**Theorem 3.** *The optimal linear matrix $\Phi^*$ for the ITM for the Gaussian measurement model in (1) under the first signal model and energy constraint in (4), for arbitrary source statistics $P_Y$, is*

$$\Phi^* = U_\Phi^* D_\Phi^* V_\Phi^{*T}, \tag{10}$$

*where $U_\Phi^* = U_\Sigma$, $V_\Phi^* = U_{E'_\beta}$ and $D_\Phi^* = [\mathrm{diag}(\sqrt{\lambda_1^*},\ldots,\sqrt{\lambda_m^*}), \mathbf{0}]$. Define*

$$f_i(\beta,\lambda_i^*) = \sigma_i^2\{[U_{E'_\beta}^T \tilde{E} U_{E'_\beta}]_{ii} - \beta[U_{E'_\beta}^T E U_{E'_\beta}]_{ii}\}^{-1} \tag{11}$$

*The elements $\lambda_1^*, \ldots, \lambda_m^*$ satisfy the following fixed-point solution:*

$$\begin{aligned} &\text{if} \quad 1/\eta < f_i(\beta, 0), \quad \textit{then} \quad \lambda_i^* = 0, \\ &\textit{otherwise} \quad \lambda_i^* \ \textit{is such that} \quad 1/\eta = f_i(\beta, \lambda_i^*), \end{aligned} \tag{12}$$

*where $\eta \geq 0$ is the Lagrange multiplier which ensures the energy constraints $\frac{1}{m}\text{tr}(\Phi\Phi^T) \leq 1$. In particular, if $\beta > 1$, $\lambda_i^* = 0$ for $i = 1, \ldots, m$.*

The ITM with the orthonormality constraint under arbitrary $P_{Y|X}$ can be viewed as a optimization problem related to calculating volumes on the Stiefel manifold, where the derivation of an analytical solution is known to be particularly challenging [25]. Rather than seeking the solution under the first signal model, we alternatively address the problem with the second signal model in which $(X', Y)$ are jointly Gaussian, and the noise precision matrix is diagonal $\Lambda = \sigma^{-1} I_m$ with $\sigma > 0$. For this case a structural fixed-point type solution is presented, and analytical solutions when $\beta = 0$ and $\beta \to -\infty$ can be readily deduced. We now present a result for the solution of the ITM under the orthonormality constraint.

**Theorem 4.** *Assume the second signal model in which $(X', Y)$ are jointly Gaussian and the noise precision $\Lambda = \sigma^{-1} I_m$ with $\sigma > 0$. There exists a nonsingular matrix $W(\Phi) \in \mathbb{R}^{m \times m}$ which simultaneously diagonalizes $\Phi(\Sigma_Y + \sigma I_m)\Phi^T$ and $\Phi(\Sigma_{Y|X'} + \sigma I_m)\Phi^T$, such that $W\Phi(\Sigma_Y + \sigma I_m)\Phi^T W^T = I_m$ and $W\Phi(\Sigma_{Y|X'} + \sigma I_m)\Phi^T W^T = \text{diag}\{\alpha_1, \ldots, \alpha_m\}$. We have that $\{\alpha_i\}_{i=1}^m$ is a collection of $m$ arbitrary eigenvalues out of $n$ eigenvalues $\{\omega_i\}_{i=1}^n$ of the matrix $(\Sigma_Y + \sigma I_m)^{-\frac{1}{2}}(\Sigma_{Y|X'} + \sigma I_m)(\Sigma_Y + \sigma I_m)^{-\frac{1}{2}}$, whose specific choice of those $m$ eigenvalues depends on $\Phi$., i.e., $\{\alpha_i\}_{i=1}^m = \cup_{i \in \Gamma_1(\Phi)}\{\omega_i\}$ for some $\Gamma_1(\Phi) \subset \{1, 2, \ldots, n\}$ depending on $\Phi$ with $|\Gamma_1| = m$. Let $U(\Phi) \in \mathbb{R}^{n \times m}$ be constituted by $m$ eigenvectors corresponding respectively to the eigenvalues $\{\alpha_i\}_{i=1}^m$. The optimal orthogonal linear matrix $\Phi^*$ under the orthonormality constraint with $\beta \in (-\infty, 0)$ satisfies the following fixed-point equation:*

$$\Phi^* = W^{-1}(\Phi^*)U^T(\Phi^*)(\Sigma_Y + \sigma I_m)^{-\frac{1}{2}}, \tag{13}$$

*and the following optimization:*

$$\arg\min_{\Phi}\{\log\, \det((W(\Phi))^2) - \frac{1}{\beta}\sum_{i \in \Gamma_1(\Phi)} \log \omega_i\}. \tag{14}$$

*When $\beta = 0$, the above fixed-point solution admits an analytic solution that*

$$\Phi^* = \text{orth}(U_1^T(\Sigma_Y + \sigma I_m)^{-\frac{1}{2}}), \tag{15}$$

*where $U_1$ is constituted by the eigenvectors corresponding to the smallest $m$ eigenvalues of the matrix $(\Sigma_Y + \sigma I_m)^{-\frac{1}{2}}(\Sigma_{Y|X'} + \sigma I_m)(\Sigma_Y + \sigma I_m)^{-\frac{1}{2}}$. $orth(\cdot)$ denotes an orthonormal basis of the row space of the argument matrix.*

*When $\beta \to -\infty$, the above fixed-point solution also admits an analytic solution that*

$$\Phi^* = \text{orth}(U_2^T(\Sigma_Y + \sigma I_m)^{-\frac{1}{2}}), \tag{16}$$

*where $U_2$ is constituted by the eigenvectors corresponding to the smallest $m$ eigenvalues of the matrix $(\Sigma_Y + \sigma I_m)^{-1}$.*

### 4.1 Discussion

Theorems 3 and 4 provide fixed-point type solutions for the *form* of $\Phi^*$, under energy and orthonormality constraints, respectively. Note that $f_i(\beta, \lambda_i^*)$ defined in (11) is a function of $\Phi^*$, and hence so are *all* $\{\lambda_i^*\}$ and *all* $\{\sigma_i\}$, through the matrices $E$ and $\tilde{E}$, which are both a function of the complete $\Phi^*$. Therefore, $f_i(\beta, \lambda_i^*)$ should be viewed in (11) as a function of $\lambda_i^*$, with all other parameters associated with $\Phi^*$ at their optimal settings (the $\{\sigma_i\}$ are not optimized, but are assumed known characteristics of the measurement noise). This suggests an iterative solution for the parameters of $\Phi^*$ in (10) under the energy constraint. Similarly, an iterative solution is also available for $\Phi^*$ under the orthonormality constraint via (13). In the below experiments we perform design of $\Phi^*$ based on gradient descent (using Theorems 1 and 2), since that approach is also applicable to the Poisson case we consider in experiments. The key contribution of Theorem 3 is that it provides insight into the characteristics of the solution of the ITM for Gaussian measurements but arbitrary source statistics $P_{Y|X}$; these insights are consistent with findings of experimental results in Section 5. Theorem 4 sheds light on the solution structure of the ITM under the orthonormality constraint, and provides analytic solutions for two extreme cases where $\beta = 0$ and $\beta \to -\infty$.

The values $\{f_i(\beta, 0)\}$ correspond to "steps", and if the "water level" $1/\eta$ is lower than the step level $f_i(\beta, 0)$, then the corresponding $\lambda_i^* = 0$. When the water level $1/\eta > f_i(\beta, 0)$, the corresponding $\lambda_i^* \neq 0$. The non-zero values $\lambda_i^*$ are *not* the difference $1/\eta - f_i(\beta, 0)$, but rather the value $\lambda_i^\dagger$ for which we achieve equality $1/\eta = f_i(\beta, \lambda_i^\dagger)$.

The results in Theorems 3 and 4 reduce to published results for special cases. For example, when $\beta = 0$ the problem reduces to the goal of classification based on $\tilde{Y}$, without any concern for recovering $Y$, or on the complexity of $\Phi^*$ [4]. In this case $f_i(\beta = 0, \lambda_i^*)$ is only a function of the generalized MMSE matrix $\tilde{E}$, and does not depend on the spectral properties of $E$, which is explicitly tied to reconstruction quality. It can also be demonstrated that in the limiting case where $\beta \to -\infty$, we recover the result in [8], which was only concerned with recovering $Y$ from $\tilde{Y}$.

We now examine what the result in Theorem 3 implies. First consider the role of $\sigma_i$ in $f_i(\beta, \lambda_i^*)$. The smaller the noise variance $\sigma_i^2$, the more likely $1/\eta > f_i(\beta, 0)$, and therefore the more likely that the $i$th eigenvector of the noise covariance will contribute to utilized $U_\Phi^*$ on the left decomposition of $\Phi^*$. What this says is that $\Phi^* Y \in \mathbb{R}^m$ resides in a linear subspace of $\mathbb{R}^m$ spanned by the weakest-energy noise eigenvectors (the number of which are "turned on" depends on the water level $1/\eta$ which is connected to the energy constraint). This is expected, as by utilizing the eigenvectors of $\Sigma$ where the additive noise is weakest to define utilized elements of $U_\Phi^*$, $\tilde{Y}$ resides in a linear subspace where the noise is weakest. This characteristic of the left singular vectors of $\Phi^*$ is consistent

across all designs of for a Gaussian measurement model [4], [8].

Concerning the right singular vectors of $\Phi^*$, they are linked to the eigenspectrum of the MMSE matrix $E$ and the generalized form $\tilde{E}$. For the simplified case of a Gaussian source $P_Y$ and a focus only on recovery $\beta \to -\infty$, this simplifies to the case in [8] for which the right singular vectors of $\Phi^*$ span a linear subspace where the *input signal* is largest, corresponding to the principal components of the source.

We now consider the role of $\beta$ under the energy constraint, which from (3) impacts the care one places in recovery (the degree to which $I(Y;\tilde{Y})$ is emphasized). For large and positive $\beta$, large values of $I(Y;\tilde{Y})$ are de-emphasized, implying that the measurement matrix $\Phi^*$ should only capture the most "relevant" information about the label $X$, but the "complexity" of $\Phi^*$ should be as little as possible. This was discussed in [19] heuristically for a Gaussian source, where here we extend to arbitrary input statistics $P_{Y|X}$, and we make the heuristic reasoning explicit. Note that as $\beta$ becomes more positive, more indices $i$ are likely to satisfy the inequality $\eta > \frac{1}{\sigma_i^2}\{[U_{E'_\beta}^T \tilde{E} U_{E'_\beta}]_{ii} - \beta[U_{E'_\beta}^T E U_{E'_\beta}]_{ii}\}$, which means that the associated $\lambda_i^* = 0$. This demonstrates explicitly that with increasing $\beta$ the notion of reducing the complexity of $\Phi^*$ means that its rank is diminished; we emphasize and demonstrate below that this defines the number of needed measurements $m$. A reduced rank of $\Phi^*$ implies that fewer measurements (features) are used, because the number of measurements may be set to the rank. Hence, for increasing $\beta > 0$ the "bottleneck" in the ITM is the number of features/measurements used for classification.

While Theorem 3 is only for a Gaussian measurement model, we observed the same phenomenon with respect to $\beta$ for the Poisson measurement model. Specifically, we observed that increasing $\beta$ favors lower-rank $\Phi^*$, even in the Poisson case (detailed in Section 5.4).

According to Theorem 4, the optimal orthogonal projection is applied to the input signal $Y$ via the following three steps. First, $Y$ is multiplied by a covariance matrix $(\Sigma_Y + \sigma I_m)^{-\frac{1}{2}}$, and signal $Y$ is transformed in the way that the variance of noise $\Sigma$ is incorporated. The second step, where an essential compression occurs, is that the transformed signal is then mapped via $U^T$ to the eigenspaces spanned by $m$ eigenvectors of $(\Sigma_Y + \sigma I_m)^{-\frac{1}{2}}(\Sigma_{Y|X'} + \sigma I_m)(\Sigma_Y + \sigma I_m)^{-\frac{1}{2}}$, and the specific choice of such $m$ eigenvectors is manifested via the optimization in (14). Finally, an invertible transform $W^{-1}$ is applied to form an orthonormal basis of the compressed $m$-dimensional space.

Since $\Phi$ is naturally required to be full-rank, the range of $\beta$ under the orthonormality constraint is confined as $\beta \in (-\infty, 0]$. When $\beta \to -\infty$, the ITM is simplified to the case where pure signal recovery is considered, and $\beta = 0$ corresponds to the pure classification problem. As we observe from the analytic expressions presented in Theorem 4, in the former extreme case, the $m$-dimensional eigenmatrix $U_2$ solely depends on the signal variance $\Sigma_Y$ itself, whereas in the latter case, the conditional variance $\Sigma_{Y|X'}$ plays a role in the matrix $U_1$.

# 5 Experiments

## 5.1 Summary of the Proposed Algorithms

The experiments considered below are performed in the context of the first model discussed in Section 2.1 (discrete $X$). There are three steps associated with implementing the above methods in the subsequent experiments:

1) Learn $P_{Y|X}$ and $P_Y$, based on available training data. The $X$ corresponds to the class label in the examples considered. The distribution $P_{Y|X}$ is modeled as a GMM, and the GMM parameters are learned using the EM algorithm based on training data, as in [36]. There is a discrete probability distribution on $X$ (probability of each data class), and therefore $P_Y$ is also a GMM (discussed further below).
2) Given the learned signal statistics, design the optimal projection $\Phi$ via the algorithm described in Section 3.2.
3) In the testing phase, we test the accuracy of the estimated $\hat{X}$ and/or $\hat{Y}$, based on the designed $\Phi$. To estimate $\hat{X}$ and/or $\hat{Y}$, we use the GMM signal model from Step 1. To estimate $\hat{Y}$, we use the same analytic CS inversion method as introduced in [36]. When estimating $\hat{X}$, we maximize $P_{X|Y}$ (which can be expressed analytically, given the signal model).

Providing further details, in our experiments for the Gaussian measurement model, we impose the GMM $P_{Y|X=i} = \sum_{j=1}^{L^{(i)}} \nu_j^{(i)} \mathcal{N}(\mu_j^{(i)}, \Sigma_j^{(i)})$, meaning that the distribution of $Y$ for each discrete $X$ is a GMM. Further, $P_Y$, with the class label summed out is also a GMM: $P_Y = \sum_{i=1}^{L} P_{Y|X=i} P_{X=i} = \sum_{i=1}^{L} \sum_{j=1}^{L^{(i)}} \pi_i \nu_j^{(i)} \mathcal{N}(\mu_j^{(i)}, \Sigma_j^{(i)})$, where $\nu_j^{(i)}$, $j = 1, \ldots, L^{(i)}$ are the GMM coefficients within class $i$. $\mu_j^{(i)}$ and $\Sigma_j^{(i)}$, $j = 1, \ldots, L^{(i)}$ are the means and covariance matrices of the respective $L^{(i)}$ Gaussian components. As detailed in [4], [36], the GMM parameters may be readily learned based on training data.

In addition to being easily sampled, the GMM has the advantage of an analytic posterior $P_{Y|\tilde{Y}}$, which is also a GMM [36], specifically $P_{Y|\tilde{Y}} = \sum_{i=1}^{L} \tilde{\pi}_i P_{Y|\tilde{Y},X=i}$, with analytic expressions for $\{\tilde{\pi}_i\}$ and $P_{Y|\tilde{Y},X=i}$ (see [36] for details on these GMM expressions). This posterior naturally induces an analytical Bayesian classifier $\max_i P_{X=i|\tilde{Y}}$, where $P_{X=i|\tilde{Y}} = \tilde{\pi}_i$, and an analytical MMSE estimator $\hat{Y}$ for the input signal $Y$, $\hat{Y} := \mathbb{E}[Y|\tilde{Y}] = \sum_{i=1}^{L} \tilde{\pi}_i \sum_{j=1}^{L^{(i)}} \tilde{\nu}_j^{(i)} \tilde{\mu}_j^{(i)}$.

## 5.2 Related Methods

IDA [12] and LDA [26] are two popular information-theoretic supervised dimensionality reduction algorithms, and they both are derived for a mixture model for $Y$, across the $L$ classes, i.e., $P_Y = \sum_{i=1}^{L} \pi_i \mathcal{N}(\mu_i, \Sigma_i)$ (LDA and IDA assume a *single* Gaussian per class). LDA aims to simultaneously maximize between-class features and minimize the scatter of the projected data within each class. Let $\mu_Y = \sum_{i=1}^{L} \pi_i \mu_i$ denote the mean of $Y$, $\Sigma_Y = \sum_{i=1}^{L} \pi_i (\Sigma_i + (\mu_i - \mu_Y)(\mu_i - \mu_Y)^T)$ represents the covariance of $Y$, and $\Lambda$ is the precision matrix of the Gaussian noise. The information-theoretic criterion $I_{LDA}(X;\tilde{Y})$ is

TABLE 1
Classification Accuracy on the Satellite Datasets under the Energy Constraint with $L^{(i)} = 5$

| $m$ | IDA | LDA | Rényi | PC | PSR | $\beta = -10$ | $\beta = -5$ | $\beta = -2$ | $\beta = -1.5$ | $\beta = -1$ | $\beta = -0.5$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.4940 | 0.6325 | 0.6213 | **0.6455** | 0.4835 | 0.4815 | 0.4840 | 0.4860 | 0.4865 | 0.4890 | 0.4935 |
| 2 | 0.8190 | 0.7925 | 0.8012 | **0.8295** | 0.8230 | 0.8270 | 0.8265 | 0.8275 | 0.8260 | 0.8262 | 0.8265 |
| 3 | 0.8565 | 0.8535 | 0.8551 | **0.8600** | 0.8400 | 0.8415 | 0.8440 | 0.8450 | 0.8545 | 0.8550 | 0.8575 |
| 4 | 0.8675 | 0.8560 | 0.8610 | **0.8695** | 0.8525 | 0.8545 | 0.8620 | 0.8625 | 0.8620 | 0.8675 | 0.8690 |
| 5 | 0.8685 | 0.8640 | 0.8655 | 0.8680 | 0.8650 | 0.8660 | 0.8675 | 0.8685 | 0.8690 | 0.8685 | **0.8695** |

*PC and PSR stand for pure classification (as in [4]) and pure signal recovery (as in [8]), which correspond to the case $\beta = 0$ and $\beta = -1{,}000$, respectively.*

$$I_{LDA}(X;\tilde{Y}) = \frac{1}{2}\log\left((2\pi e)^m \det(\Phi\Sigma_Y\Phi^T + \Lambda^{-1})\right) - \frac{1}{2}\log\left((2\pi e)^m \det\left(\Phi\left(\sum_{i=1}^{L}\pi_i\Sigma_i\right)\Phi^T + \Lambda^{-1}\right)\right), \tag{17}$$

The optimal solution maximizing $I_{LDA}(X;\tilde{Y})$ is available analytically [26].

IDA seeks to maximize mutual information $I(X;\tilde{Y})$ directly, which can be expressed as $I_{IDA}(X;\tilde{Y}) = h(\tilde{Y}) - h(\tilde{Y}|X)$, where $h(\cdot)$ is the differential entropy, and $h(\tilde{Y}|X) = \frac{1}{2}\sum_{i=1}^{L}\pi_i \log\left((2\pi e)^m \det(\Phi\Sigma_i\Phi^T + \Lambda^{-1})\right)$. In order to calculate $h(\tilde{Y})$, IDA imposes a Gaussian approximation, with learned covariance $\Sigma_Y$. Hence, under this approximation, $h(\tilde{Y}) = \frac{1}{2}\log\left((2\pi e)^m \det(\Phi\Sigma_Y\Phi^T + \Lambda^{-1})\right)$. The maximization on $I_{IDA}(X;\tilde{Y})$ can be carried out by the gradient descent method [12].

Making a connection to the proposed ITM framework in (3), consider the case $\beta = 0$, for which we are only interested in classification. Further, assume only a single Gaussian (not a GMM) per class $X$. In this case the ITM reduces to the IDA criterion. Hence, in the limit $\beta \to 0$, the difference between our ITM and IDA is that in ITM we employ a GMM per class $X$. Of course, the ITM also extends to cases $\beta \neq 0$.

In order to apply LDA and IDA, we must convert the model $P_{Y|X=i} = \sum_{j=1}^{L^{(i)}} \nu_j^{(i)} \mathcal{N}(\mu_j^{(i)}, \Sigma_j^{(i)})$ used in our projection design to a form that LDA and IDA may employ (single Gaussian per class). We use the overall mean and covariance for class $X = i$ as the associated parameters of the class-$i$ Gaussian model. While this simplified model is used in LDA and IDA *design* , classification and signal recovery with the $\Phi$ so learned are performed using the full mixture model, with $P_{Y|X=i} = \sum_{j=1}^{L^{(i)}} \nu_j^{(i)} \mathcal{N}(\mu_j^{(i)}, \Sigma_j^{(i)})$. Therefore, all algorithms studied here use the same model for classification and for estimation of $Y$, and the only difference is manifested in how $\Phi$ is learned. A similar LDA and IDA design and test procedure was employed in [4].

We also consider the information-theoretic supervised dimensionality reduction method proposed in [2], [3]. Instead of using the Shannon entropy, they used quadratic Rényi entropy to define the mutual information as

$$I_T(X;\tilde{Y}) = \sum_{i=1}^{L}\int (P_{\tilde{Y}=y,X=i} - P_{\tilde{Y}=y}P_{X=i})^2 dy. \tag{18}$$

The derivative of the quadratic Rényi mutual information can be expressed analytically for the GMM signal [2], [3].

### 5.3 Satellite and USPS Data

For the Gaussian measurement model in (1), we conduct experiments on the satellite and USPS datasets. Both datasets are available online and have been used in [4], [12]. The satellite dataset contains 36-dimensional feature vectors of satellite data, comprised of pixel values of a $3 \times 3$ neighborhood in 4 spectral channels. The six class labels for the central pixel are real soil, cotton crop, grey soil, damp grey soil, soil with vegetation stubble, and very damp grey soil. The training set contains 4,435 samples, and the testing set contains 2,000 samples. The USPS data contains grey scale images of dimension 256 for handwritten ten single-digits. There are 7,291 training samples and 2,007 testing samples.

As stated above, the parameters of the GMM are learned via the EM algorithm. Unless otherwise stated, we consider $L^{(i)} = 5$ mixture components for each class; similar results were found for $L^{(i)} = 1$ and $L^{(i)} = 10$. The noise covariance matrix $\Sigma$ is set to be $10^{-6}I_m$, where $I_m$ is the $m \times m$ identity matrix (we are essentially doing feature learning in this application, as in [4]). When performing the Monte Carlo integrations, 500 samples were applied to calculate $E$ and $\tilde{E}$. The performance of the gradient descent algorithm can be significantly affected by the choice of step size, and smaller step size usually leads to a better performance but a slower convergence rate. For these two datasets, we find that a good trade-off between speed and performance is achieved when the step size is chosen between 0.1 to 1 percent of the $\mathrm{tr}(\Phi\Phi^T)$, and we use 500 gradient iterations in the experiments on these two datasets. We compare performance of the proposed ITM under various settings of $\beta$ to the LDA, IDA and Rényi methods.

In Tables 1, 2, 3 and 4, we present respectively classification accuracy and fractional error for signal recovery on the Satellite data under the energy and orthonormality constraints with $L^{(i)} = 5$, where the fractional error is defined as $\frac{\|\hat{Y}-Y\|_2^2}{\|Y\|_2^2}$, where $\hat{Y}$ is the estimate of $Y$. As explained in the previous section, the pure classification and pure signal recovery cases correspond to $\beta = 0$ and $\beta = -1{,}000$, respectively (values of $\beta < -1{,}000$ yielded almost identical results, indicating that this approached the asymptotic limit).

We also present classification accuracy and fractional error for signal recovery on the Satellite data under the energy constraint with $L^{(i)} = 1$ in Tables 5 and 6. The results under orthonormality constraint are similar, and are omitted for brevity.

Under both the energy and orthonormality constraints, it can be easily observed that the advantage of the ITM for joint classification and recovery comes from $\beta < 0$, when

TABLE 2
Fractional Error of the Signal Recovery on the Satellite Datasets with Various Values of $\beta$ under the Energy Constraint $L^{(i)} = 5$

| $m$ | IDA | LDA | Rényi | PC | PSR | $\beta = -10$ | $\beta = -5$ | $\beta = -2$ | $\beta = -1.5$ | $\beta = -1$ | $\beta = -0.5$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.2632 | 2.5170 | 2.3270 | 2.3882 | **2.2552** | 2.2588 | 2.2608 | 2.2704 | 2.2761 | 2.2903 | 2.3334 |
| 2 | 0.6451 | 1.9027 | 0.6761 | 0.5558 | **0.5337** | 0.5341 | 0.5377 | 0.5430 | 0.5449 | 0.5482 | 0.5524 |
| 3 | 0.6244 | 0.6876 | 0.6512 | 0.5233 | 0.4250 | **0.4216** | 0.4311 | 0.4400 | 0.4501 | 0.4604 | 0.4897 |
| 4 | 0.5143 | 0.5975 | 0.5170 | 0.4718 | **0.3204** | 0.3271 | 0.3275 | 0.3285 | 0.3290 | 0.3298 | 0.3320 |
| 5 | 0.4871 | 0.5813 | 0.5042 | 0.3889 | **0.2903** | 0.2909 | 0.2919 | 0.2930 | 0.2929 | 0.2913 | 0.2989 |

TABLE 3
Classification Accuracy on the Satellite Datasets under the Orthonormality Constraint with $L^{(i)} = 5$

| $m$ | IDA | LDA | Rényi | PC | PSR | $\beta = -10$ | $\beta = -5$ | $\beta = -2$ | $\beta = -1.5$ | $\beta = -1$ | $\beta = -0.5$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.4840 | 0.5385 | 0.5211 | **0.5940** | 0.4785 | 0.4805 | 0.4810 | 0.4835 | 0.4860 | 0.4935 | 0.4900 |
| 2 | 0.8240 | 0.7760 | 0.7904 | **0.8290** | 0.7990 | 0.8050 | 0.8070 | 0.8080 | 0.8095 | 0.8125 | 0.8255 |
| 3 | 0.8480 | 0.8500 | 0.8502 | **0.8525** | 0.8235 | 0.8280 | 0.8300 | 0.8315 | 0.8420 | 0.8465 | 0.8480 |
| 4 | 0.8680 | 0.8575 | 0.8619 | **0.8710** | 0.8610 | 0.8615 | 0.8620 | 0.8635 | 0.8640 | 0.8655 | 0.8695 |
| 5 | 0.8720 | 0.8620 | 0.8683 | 0.8725 | 0.8670 | 0.8635 | 0.8645 | 0.8675 | 0.8675 | 0.8715 | **0.8795** |

*PC and PSR stand for pure classification (as in [4]) and pure signal recovery (as in [8]), which correspond to the case $\beta = 0$ and $\beta = -1{,}000$, respectively.*

TABLE 4
Fractional Error of Signal Recovery on the Satellite Datasets with Various Values of $\beta$ under the Orthonormality Constraint with $L^{(i)} = 5$

| $m$ | IDA | LDA | Rényi | PC | PSR | $\beta = -10$ | $\beta = -5$ | $\beta = -2$ | $\beta = -1.5$ | $\beta = -1$ | $\beta = -0.5$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.3318 | 3.7341 | 2.8723 | 2.4284 | **2.2841** | 2.2847 | 2.2852 | 2.2896 | 2.2949 | 2.3071 | 2.3356 |
| 2 | 0.6811 | 1.4982 | 0.7497 | 0.6846 | **0.6240** | 0.6286 | 0.6365 | 0.6439 | 0.6510 | 0.6684 | 0.6773 |
| 3 | 0.6524 | 0.6111 | 0.6237 | 0.5499 | **0.4906** | 0.4976 | 0.4993 | 0.5004 | 0.4993 | 0.5203 | 0.5287 |
| 4 | 0.5343 | 0.5393 | 0.5562 | 0.4504 | **0.3617** | 0.3633 | 0.3632 | 0.3782 | 0.3797 | 0.3863 | 0.3894 |
| 5 | 0.4284 | 0.5228 | 0.4342 | 0.3883 | 0.2906 | **0.2901** | 0.2909 | 0.2987 | 0.2983 | 0.3006 | 0.3079 |

TABLE 5
Classification Accuracy on the Satellite Datasets under the Energy Constraint with $L^{(i)} = 1$

| $m$ | IDA | LDA | Rényi | PC | PSR | $\beta = -10$ | $\beta = -5$ | $\beta = -2$ | $\beta = -1.5$ | $\beta = -1$ | $\beta = -0.5$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.4515 | 0.5822 | 0.5835 | **0.6315** | 0.4592 | 0.4614 | 0.4620 | 0.4625 | 0.4633 | 0.4645 | 0.4690 |
| 2 | 0.7918 | 0.7714 | 0.7892 | **0.8083** | 0.7995 | 0.8014 | 0.8025 | 0.8033 | 0.8045 | 0.8053 | 0.8071 |
| 3 | 0.8218 | 0.8227 | 0.8220 | **0.8412** | 0.8294 | 0.8304 | 0.8314 | 0.8325 | 0.8333 | 0.8345 | 0.8352 |
| 4 | 0.8282 | 0.8301 | 0.8293 | **0.8514** | 0.8383 | 0.8403 | 0.8416 | 0.8422 | 0.8438 | 0.8450 | 0.8463 |
| 5 | 0.8323 | 0.8338 | 0.8319 | **0.8592** | 0.8321 | 0.8330 | 0.8346 | 0.8359 | 0.8366 | 0.8375 | 0.8389 |

*PC and PSR stand for pure classification (as in [4]) and pure signal recovery (as in [8]), which correspond to the case $\beta = 0$ and $\beta = -1{,}000$, respectively.*

TABLE 6
Fractional Error of the Signal Recovery on the Satellite Datasets with Various Values of $\beta$ under the Energy Constraint with $L^{(i)} = 1$

| $m$ | IDA | LDA | Rényi | PC | PSR | $\beta = -10$ | $\beta = -5$ | $\beta = -2$ | $\beta = -1.5$ | $\beta = -1$ | $\beta = -0.5$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.3812 | 2.4131 | 2.3911 | 2.3982 | **2.3452** | 2.3481 | 2.3542 | 2.3601 | 2.3663 | 2.3722 | 2.3783 |
| 2 | 0.6812 | 2.1396 | 0.7821 | 0.5912 | **0.5512** | 0.5541 | 0.5617 | 0.5680 | 0.5712 | 0.5753 | 0.5812 |
| 3 | 0.6512 | 0.7123 | 0.7029 | 0.5612 | **0.4325** | 0.4416 | 0.4443 | 0.4480 | 0.4513 | 0.4652 | 0.4712 |
| 4 | 0.6125 | 0.6240 | 0.6284 | 0.5016 | **0.3641** | 0.3671 | 0.3718 | 0.3807 | 0.3893 | 0.3984 | 0.4012 |
| 5 | 0.5471 | 0.5618 | 0.5345 | 0.4234 | **0.3121** | 0.3292 | 0.3378 | 0.3406 | 0.3484 | 0.3581 | 0.3691 |

one can balance the goals of signal classification (Tables 1, 3 and 5) and recovery (Tables 2, 4 and 6). For $m = 1$, it seems that the classification performance is less sensitive to the tuning of $\beta$. Note that for $\beta = -0.5$ and for number of measurements $m \geq 2$, the ITM classification accuracy is almost as good as the pure-classification case ($\beta = 0$ and [4]), and better than LDA, IDA and Rényi method). However, there is a marked gain in recovery accuracy versus these three alternatives, as indicated in Tables 2, 4 and 6. Such a designed $\Phi$ has the significant advantage of excellent recovery and classification simultaneously.

When $\beta > 0$, ITM considers pure classification, and increasing $\beta$ further penalizes signal recovery ($\beta > 0$ seeks the fewest number of classification features). More specifically, the rank of the projection matrix is controlled by the value of $\beta > 0$, as suggested by Theorem 3. In order to

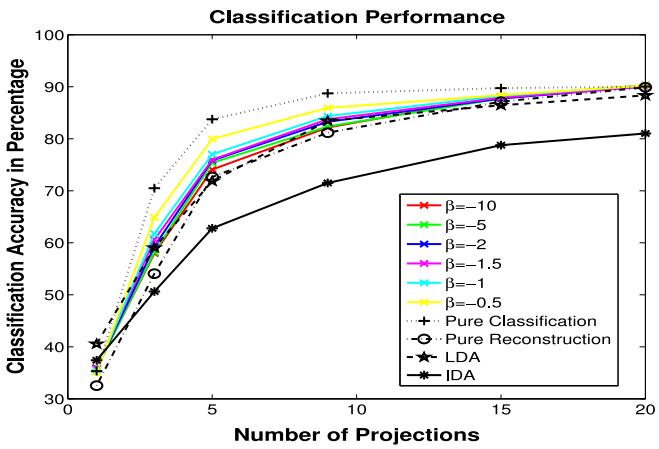
TABLE 7
Classification and Rank under Various Values of $\beta$ on the Satellite Datasets

| $\beta$ | Rank | CA | CA (rank-reduced) |
|---|---|---|---|
| 100 | 1 | 0.6343 | 0.6311 |
| 50 | 2 | 0.8139 | 0.8083 |
| 10 | 4 | 0.8642 | 0.8622 |
| 1 | 9 | 0.9012 | 0.8981 |

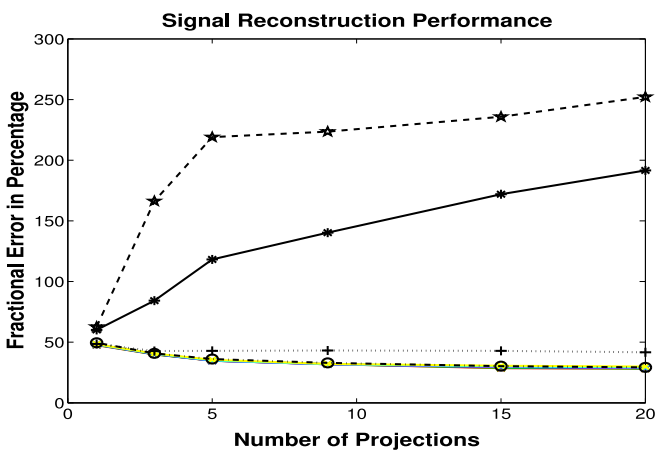
*CA stands for classification accuracy.*

verify the theoretical result, we set $m = n$, i.e., $\Phi \in \mathbb{R}^{n\times n}$. Note that in our experiments, $\Phi$ is non-zero for any $\beta$, since we implement the energy constraint as $\frac{1}{m}\mathrm{tr}(\Phi\Phi^T) = 1$ in all the experiments.

In order to calculate an approximate rank, we threshold the singular values of the numerically obtained projection matrix at $10$ percent of the largest singular value (because of the numerics, $\Phi^*$ is not exactly low-rank, as stipulated in Theorem 3). In order to obtain the rank-reduced $\Phi^* \in \mathbb{R}^{r\times n}$, one may truncate $\Phi^*$ to be in $\mathbb{R}^{r\times n}$, by taking any $r$ linearly independent rows and normalizing to satisfy the energy constraint on $\mathrm{trace}(\Phi^*(\Phi^*)^T)$. In

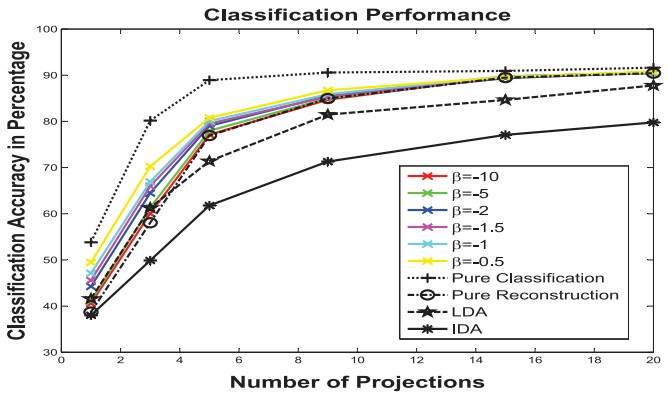



(a)

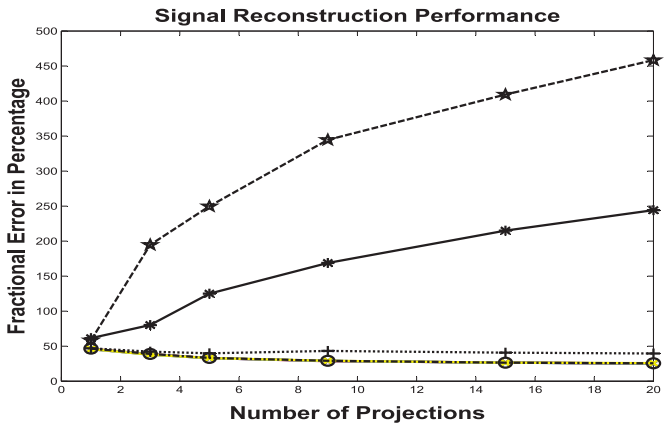



(b)

Fig. 1. Classification accuracy and fractional error for signal recovery on the USPS datasets under the energy constaint. (a) The classification accuracy under various values of $\beta$. (b) The fractional error of signal recovery under various values of $\beta$. In (b), the same symbol identifications are used as in (a).




(a)




(b)

Fig. 2. Classification accuracy and fractional error for signal recovery on the USPS datasets under the orthonormality constraint. (a) The classification accuracy under various values of $\beta$. (b) The fractional error of signal recovery under various values of $\beta$. In (b), the same symbol identifications are used as in (a).

Table 7, we list the rank and classification performances under various values of $\beta$.

The rank-controlling property of the ITM, for $\beta > 0$, has important practical importance and $\beta > 0$ defines the number of measurements needed, an important output of the ITM setup. The performance of the original $\Phi^* \in \mathbb{R}^{n\times n}$ and the rank-reduced and renormalized $\Phi^* \in \mathbb{R}^{r\times n}$ were virtually identical, for signal classification.

We next consider the USPS dataset. In Figs. 1 and 2 we show signal classification and recovery results under the energy and orthonormality constraints, for the same cases as considered in the above tabular results, as a function of $\beta < 0$, but for a larger range of measurements $m$. We observe that for classification under both constraints, quantified in Figs. 1a and 2a, IDA performs poorly and the performance of ITM gradually grows with increase of $\beta$, with the best result achieved under $\beta = 0$ (which corresponds to [4]). However, the $\beta = 0$ case yields poor signal recovery for both cases (Figs. 1b and 2b), and LDA and IDA yield particularly poor recovery results. The results for the Rényi method are similar to IDA, and are omitted for brevity. Part of the reason for the poor performance of IDA in the classification case may be due to the fact that IDA approximates the mutual information $I(X;\tilde{Y})$ up to second-order statistics, which does not work well on the USPS datasets. Again, for

TABLE 8
Classification and Rank under Various Values of $\beta$ on the USPS Datasets

| $\beta$ | Rank | CA | CA (rank-reduced) |
|---|---|---|---|
| 500 | 1 | 0.3416 | 0.3410 |
| 400 | 2 | 0.4129 | 0.4092 |
| 100 | 6 | 0.8812 | 0.8821 |
| 50 | 11 | 0.9131 | 0.9119 |

*CA stands for classification accuracy.*

$\beta = -0.5$, the ITM setup yields a good balance between recovery and classification.

For the USPS data, we consider $\beta > 0$ in Table 8. We again observe the rank-controlling property of $\beta > 0$, this defining the number of measurements/features for classification, with recovery penalized.

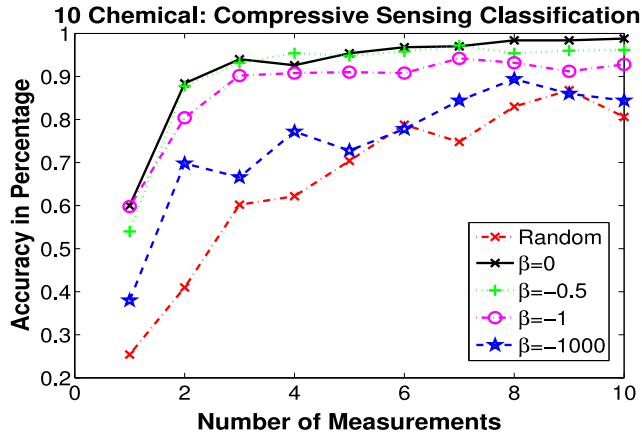
### 5.4 Poisson Compressive Hyperspectral Camera

We consider a compressive chemical sensing system based on the wavelength-dependent signature of chemicals, at optical frequencies. The details of the measurement system are provided in [27]. Summarizing briefly, multi-wavelength (hyperspectral) photons are scattered off a sample, and a digital mirror microdevice (DMD) is used to perform *binary* linear projections (binary projections on the frequency-dependent signature). For our purposes the key points are that the data are Poisson, as in (2) and the elements of the measurement matrix $\Phi$ are either 1 or 0. Neither [8] nor [4] considered the Poisson measurement model.

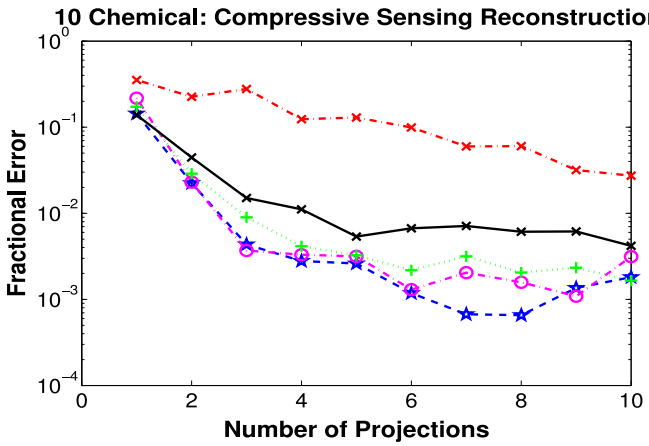
Assume that there are $T$ chemicals of interest, and that the hyperspectral sensor performs measurements at $n$ wavelengths. The observed data are represented $\tilde{Y}|X \sim \text{Pois}(\tilde{Y}; \Phi(\Psi \exp(Z_X) + \lambda))$, where $Y = \Psi \exp(Z_X)$, $Z_X \in \mathbb{R}^T$ is conditioned on label $X$, $\tilde{Y} \in \mathbb{Z}_+^n$ represents the count of photons at the $n$ sensor wavelengths, $\lambda \in \mathbb{R}_+^n$ represents the sensor dark current, and the $t$-th column of $\Psi \in \mathbb{R}_+^{n \times T}$ reflects the mean Poisson rate for chemical $t$. The expression $\exp(Z_X)$ denotes a pointwise exponentiation of $Z_X$, and $Z_X$ is drawn from a GMM. One Gaussian per chemical was sufficient.

This signal model corresponds to Poisson factor analysis [37], where $\Psi$ are the factor loadings and $\exp(Z_X)$ are the factor scores; we have also added a dark current. We here consider $T$ factors, as there are $T$ chemicals, but this is not necessary. Methods like those discussed in [37] may be used to learn the model. The measurements for this study were performed in our laboratory, and will be made available for others to experiment with.

The measurement matrix, which controls the states of the micro-mirrors, is binary, i.e., $\Phi \in \{0,1\}^{m \times n}$. Instead of directly optimizing $\Phi$, we put a logistic link on each value $\Phi_{ij} = \text{logit}(M_{ij})$. By the chain rule, we can state the gradient with respect to $M$ as: $[\nabla_M \mathcal{IB}(\Phi(M), \beta)_{ij}] = [\nabla_\Phi \mathcal{IB}(\Phi, \beta)_{ij}]$ $[\nabla_M \Phi_{ij}]$. Matrix $M$ was initialized at random, and we threshold the logistic at 0.5 to get the final binary $\Phi$. 100 Monte Carlo samples are used to calculate the MMSE matrix. 1,000 iterations of the gradient-descent have been used. Ten ($T = 10$) chemicals are considered in this test: acetone, acetonitrile, benzene, dimethylacetamide, dioxane, ethanol, hexane, methylcyclohexane, octane, and toluene. A Bayesian classifier $\max_i P_{X=i|\tilde{Y}}$ and MMSE estimator $\hat{Y} = \mathbb{E}[Y|\tilde{Y}]$ are readily employed for classification and signal recovery problems, respectively, and they can be calculated as follow.




(a)




(b)

Fig. 3. Classification accuracy and fractional error for signal recovery on the chemical sensing dataset. (a) The classification accuracy under various values of $\beta$. (b) The fractional error of signal recovery under various values of $\beta$. In (b), the same symbol identifications are used as in (a).

The posterior of the class label $X$ can be expressed as

$$P_{X=i|\tilde{Y}} = \frac{P_{\tilde{Y}|X=i} P_{X=i}}{\int_X P_{Y|\tilde{X}=i} P_X} = \frac{P_{\tilde{Y}|X=i} P_{X=i}}{\sum_{j=1}^T P_{Y|\tilde{X}=j} P_{X=j}}, \tag{19}$$

where $P_{\tilde{Y}|X=i} = \int_Z \text{Pois}(\tilde{Y}; \Phi(\Psi \exp(Z) + \lambda)) \mathcal{N}(Z; \mu_i, \Lambda_i^{-1})$, which can be readily calculated via Monte Carlo integration.

In order to calculate the posterior $P_{Y|\tilde{Y}}$, we employ the Laplace approximation [33]. More specifically, each mixture component in the prior $Z_X \sim \sum_{i=1}^T \pi_i \mathcal{N}(\mu_i, \Lambda_i^{-1})$ may be viewed as a model for the observed data, and $\pi_i$ represents the probability of model $i$. Using the Poisson likelihood function $\text{Pois}(\Phi(\Psi \exp(Z_X) + \lambda))$, we may apply a Laplace approximation for the posterior of $Z_X$ for model/mixture component $i$. Via the Laplace approximation, we update each mixture-component prior $\mathcal{N}(\mu_i, \Lambda_i^{-1})$ to a Laplace-approximated posterior. This is done for each mixture component $i \in \{1, \ldots, L\}$, and the model evidence is employed for the updated/posterior mixture probabilities. The overall

TABLE 9
Classification and Rank under Various Values of $\beta$ on the Chemical-Sensing Dataset

| $\beta$ | Rank | CA | CA (rank-reduced) |
|---|---|---|---|
| 1 | 1 | 0.6771 | 0.6766 |
| 0.9 | 4 | 0.8664 | 0.8583 |
| 0.7 | 8 | 0.9012 | 0.8931 |
| 0.5 | 9 | 0.9411 | 0.9373 |

*CA stands for classification accuracy.*

posterior on $Z_X$ is an updated Gaussian mixture model, and the MMSE estimator is the mean of the updated Gaussian mixture model.

In Fig. 3 we present the *average* performances of classification and signal recovery under various settings of $\beta$ (we learn a signal model based on measured training data, and that model is used with different $\beta$ for design of $\Phi$); 10 experiments were performed for each case, and the error bars were tight (omitted for simplity, and to not clutter the figures). The $\beta = 0$ results are best for classification (subfigure (a)) and $\beta = -1000$ is best for recovery (subfigure (b)). Random design performs poorly (probability of $1/0$ in $\Phi$ set at 0.5). As in the previous examples, by setting $\beta = -0.5$, we achieve a good balance between classification and recovery performance (here recovery of the factor score).

In Table 9, we summarize design results when $\beta > 0$ where we set $m = n$, and via the rank infer the number of needed measurements $r$ (in this case we are only interested in classification, and $\beta$ controls the number of measurements). It is observed that the rank of the projection is manipulated by the value of $\beta$, which is similar to the Gaussian case.

## 6 Conclusions

We have developed an information-theoretic metric (ITM) for linear dimensionality reduction. Gradients of the ITM have been derived for Gaussian and Poisson measurement models. Two types of constraints have been considered and the structure of the optimal linear projections have been derived explicitly for the Gaussian-measurement case. It has been shown that the proposed ITM is able to preserve simultaneously the information for signal recovery and classification when ITM parameter $\beta < 0$, and for $\beta > 0$ one can control the number of measurements for classification (number of linear features).

## Appendix A
## Appendix: Proofs of Theorems

**Proof of Theorems 1 and 2.** These two theorems directly follow from gradient results in [4], [20], [21], [32]. Note that

$$\nabla_\Phi \mathcal{ITC}(\Phi, \beta) = \nabla_\Phi I(X; \tilde{Y}) - \beta \nabla_\Phi I(Y; \tilde{Y}). \tag{20}$$

For Gaussian measurement model, by [4], we have that

$$\nabla_\Phi I(X; \tilde{Y}) = \Lambda \Phi \tilde{E}, \tag{21}$$

where

$$\tilde{E} = \mathbb{E}\Big[(\mathbb{E}[Y|\tilde{Y}, X] - \mathbb{E}[Y|\tilde{Y}])(\mathbb{E}[Y|\tilde{Y}, X] - \mathbb{E}[Y|\tilde{Y}])^T\Big].$$

By [20], [32],

$$\nabla_\Phi I(Y; \tilde{Y}) = \Lambda \Phi E, \tag{22}$$

where $E = \mathbb{E}\Big[(Y - \mathbb{E}[Y|\tilde{Y}])(Y - \mathbb{E}[Y|\tilde{Y}])^T\Big]$. Combining those two equalities, we have Theorem 1.

Similarly, for Poisson measurement model, by [21], we have that

$$\nabla_\Phi I(X; \tilde{Y}) = \mathbb{E}\left[\mathbb{E}[Y_j|\tilde{Y}, X] \log \frac{\mathbb{E}[(\Phi Y)_i|\tilde{Y}, X]}{\mathbb{E}[(\Phi Y)_i|\tilde{Y}]}\right], \tag{23}$$

and

$$\begin{aligned} &\nabla_\Phi I(Y; \tilde{Y}), \\ &= \mathbb{E}\big[Y_j \log((\Phi Y)_i)\big] - \mathbb{E}\big[\mathbb{E}[Y_j|\tilde{Y}] \log \mathbb{E}[(\Phi Y)_i|\tilde{Y}]\big]. \end{aligned} \tag{24}$$

Combining those two equalities, we have Theorem 2. □

**Proof of Theorem 3.** Consider the Karush-Kuhn-Tucker (KKT) condition of ITM, the optimal $\Phi^*$ should satisfy the follow conditions:

$$\nabla_\Phi \left[\mathcal{ITM}(\Phi, \beta) + \eta \cdot (1 - \frac{1}{m}\mathrm{tr}(\Phi\Phi^T))\right]\Big|_{\Phi=\Phi^*} = 0, \tag{25}$$

$$\eta \cdot (1 - \frac{1}{m}\mathrm{tr}(\Phi^*\Phi^{*T})) = 0, \tag{26}$$

where $\eta \geq 0$ is the Lagrange multiplier. By Theorem 1, we have

$$\frac{2}{m}\Phi^* = \Lambda \Phi^* E_\beta'^*, \tag{27}$$

$$\frac{2}{m}\Phi^*\Phi^{*T} = \Lambda(\Phi^* E_\beta'^* \Phi^{*T}). \tag{28}$$

By taking a transpose on both sides of (27), we obtain $\frac{2}{m}\Phi^{*T} = E_\beta'^*\Phi^{*T}\Lambda$, and thus $\Lambda$ and $\Phi^* E_\beta'^* \Phi^{*T}$ are commutable. By [38], two symmetric matrices are commutable if and only if they are simultaneously diagonalizable. Therefore we have

$$U_\Phi^* = U_\Lambda \Pi_U^* = U_\Sigma \Pi_U^*, \tag{29}$$

$$V_\Phi^* = U_{E_\beta'^*}^* \Pi_V^*, \tag{30}$$

where the optimal orthogonal matrices $\Pi_U^*$ and $\Pi_V^*$ leverage the effect of reordering among the associated eigenvalues. Since we have ordered the eigenvalues of $\Sigma$ and $E_\beta'^*$, without loss of generality, we may assume that $\Pi_U^* = I_m$ and $\Pi_V^* = I_n$. It is straightforward to verify that our analysis still holds with non-identity $\Pi_U^*$ and $\Pi_V^*$.

Notice that $\mathcal{ITM}(\Phi, \beta)$ and the energy constraint is invariant under arbitrary orthogonal transformation on the measurement $\tilde{Y}$. We apply the orthogonal transform $\Sigma^{1/2}U_\Sigma^T$ on $\tilde{Y}$ and obtain the equivalent Gaussian

measurement model

$$\tilde{Y}' = \Sigma^{1/2} D_\Phi V_\Phi^T Y + W', \tag{31}$$

where $W' \sim \mathcal{N}(0, I_m)$. Hence the original ITM problem is equivalent to the following optimization problem:

$$\begin{aligned} &\max_{\{\lambda_1,\ldots,\lambda_m\}} I(X;\tilde{Y}') - \beta I(Y;\tilde{Y}') \\ &\text{s.t. } \sum_i^m \lambda_i \le m \quad \text{and} \quad \lambda_i \ge 0, i = 1,\ldots,m, \end{aligned} \tag{32}$$

where $\{\lambda_1,\ldots,\lambda_m\} = \mathrm{Diag}(D) := \mathrm{Diag}(D_\Phi D_\Phi^T)$. $\mathcal{L}(D)$ is the Lagrangian

$$\begin{aligned} &\mathcal{L}(D,\eta,\eta_i) = \\ &I(X;\tilde{Y}') - \beta I(Y;\tilde{Y}') + \eta \cdot \left(m - \sum_i^m \lambda_i\right) + \sum_i^m \eta_i \lambda_i. \end{aligned} \tag{33}$$

where $\eta, \eta_i \ge 0$ are the Lagrange multipliers. By the result in [29], we have

$$\frac{\partial\Big(I(X;\tilde{Y}') - \beta I(Y;\tilde{Y}')\Big)}{\partial D} = \mathrm{Diag}(\tilde{\Lambda} D_\Sigma^{-1}), \tag{34}$$

where $\tilde{\Lambda} = U_{E'_\beta}^T E'_\beta U_{E'_\beta}$. Apply the KKT condition $\frac{\partial \mathcal{L}(D,\eta,\eta_i)}{\partial D} = \mathbf{0}$, we obtain that the optimal eigenvalues $\{\lambda_i^*\}_{i=1}^m$ should satisfy the following fixed point equation

$$\eta(D_\Sigma)_i - (\tilde{\Lambda})_i = \eta_i (D_\Sigma)_i, \tag{35}$$

where $()_i$ denotes the $i$th diagonal element and $(\tilde{\Lambda})_i$ is the $f_i$ as defined in the statement of the theorem.

By complementary slackness of KKT condition, if $\lambda_i^* > 0$, then $\eta_i = 0$. Hence, we may conclude that $\eta > 0$. Therefore, in this case, we obtain $\lambda_i^* = \lambda_i^\dagger$, where $\lambda_i^\dagger$ is a positive solution of the equation $\eta(D_\Sigma)_i = (\tilde{\Lambda})_i$. We note $(\tilde{\Lambda})_i$ is a function of $\lambda_i$.

On the other hands, if $\eta(D_\Sigma)_i = (\tilde{\Lambda})_i$ does not possess a positive solution of $\lambda_i$, then $\eta_i > 0$. Therefore, we must have $\lambda_i^* = 0$ and $\eta(D_\Sigma)_i - (\tilde{\Lambda})_i = \eta_i (D_\Sigma)_i > 0$. Thus we derive the condition $\eta(D_\Sigma)_i > (\tilde{\Lambda})_i$.

We now consider the case when $\beta > 1$. Notice that $X \to Y \to \tilde{Y}$ forms a Markov chain. By the data processing inequality [39], we have

$$I(X;\tilde{Y}) \le I(Y;\tilde{Y}). \tag{36}$$

Therefore,

$$\mathcal{ITM}(\Phi,\beta) \le (1-\beta) I(Y;\tilde{Y}). \tag{37}$$

When $\beta > 1$, we have $\mathcal{ITM}(\Phi,\beta) < 0$, given that $I(Y;\tilde{Y}) > 0$. The maximum 0 is obtained when $I(Y;\tilde{Y}) = 0$, which can be achieved by setting $\Phi = 0$, i.e., $\lambda_i^* = 0$ for all $i = 1,,m$. □

**Proof of Theorem 4.** The existence of the matrix $W$ is guaranteed by constituting $W$ via the generalized eigenvectors of $\Phi(\Sigma_{Y|X'} + \sigma I_m)\Phi^T$ and $\Phi(\Sigma_Y + \sigma I_m)\Phi^T$ [38], and such $W$ is nonsingular.

The ITM can be calculated as

$$\mathcal{ITM}(\Phi,\beta) = I(X';\tilde{Y}) - \beta I(\tilde{Y},Y), \tag{38}$$

$$= h(\tilde{Y}) - h(\tilde{Y}|X') - \beta h(\tilde{Y}) + \beta h(\tilde{Y}|Y). \tag{39}$$

Recall that for $d$-dimensional random variable $X' \sim \mathcal{N}(\mu_{X'}, \Sigma_{X'}) = \frac{1}{2}\log\left((2\pi e)^d \det(\Sigma_{X'})\right)$ [39]. Since that $(X', Y)$ is jointly Guassian, it follows that $\tilde{Y}$ and $\tilde{Y}|X'$ are also Gaussian random variables with variance $\Sigma_{\tilde{Y}} = \Phi\Sigma_Y\Phi^T + \Sigma$ and $\Sigma_{\tilde{Y}|X'} = \Phi\Sigma_{Y|X'}\Phi^T + \Sigma$, respectively [40]. Therefore, we have

$$\begin{aligned} &\arg\max_\Phi \mathcal{ITC}(\Phi,\beta), \\ &= \arg\max_\Phi \{h(\tilde{Y}) - h(\tilde{Y}|X') - \beta h(\tilde{Y}) + \beta h(\tilde{Y}|Y)\}, \end{aligned} \tag{40}$$

$$\begin{aligned} &= \arg\max_\Phi \{(1-\beta)\log\det(\Sigma_{\tilde{Y}}) - \log\det(\Sigma_{\tilde{Y}|X'}) \\ &\qquad + \beta \log\det(\Sigma)\}, \end{aligned} \tag{41}$$

$$\begin{aligned} &= \arg\max_\Phi \{\beta((\frac{1}{\beta} - 1)\log\det(\Sigma_{\tilde{Y}}) - \frac{1}{\beta}\log\det(\Sigma_{\tilde{Y}|X'}) \\ &\qquad + \log\det(\Sigma))\}, \end{aligned} \tag{42}$$

As $\beta \in (-\infty, 0)$, we have

$$\begin{aligned} &\arg\max_\Phi \mathcal{ITM}(\Phi,\beta), \\ &= \arg\min_\Phi \{(\frac{1}{\beta} - 1)\log\det(\Sigma_{\tilde{Y}}) \\ &\qquad - \frac{1}{\beta}\log\det(\Sigma_{\tilde{Y}|X'}) + \log\det(\Sigma)\}, \end{aligned} \tag{43}$$

$$\begin{aligned} &= \arg\min_\Phi \{(\frac{1}{\beta} - 1)\log\det(\Sigma_{\tilde{Y}}) \\ &\quad - \frac{1}{\beta}\log\det(\Sigma_{\tilde{Y}|X'})\}, \end{aligned} \tag{44}$$

$$\begin{aligned} &= \arg\min_\Phi \{(\frac{1}{\beta} - 1)\log\det(\Phi\Sigma_Y\Phi^T + \Sigma) \\ &\qquad - \frac{1}{\beta}\log\det(\Phi\Sigma_{Y|X'}\Phi^T + \Sigma)\}, \end{aligned} \tag{45}$$

Let $W(\Phi)$ be the matrix simultaneously diagonalizing $\Phi(\Sigma_{Y|X'} + \sigma I_m)\Phi^T$ and $\Phi(\Sigma_Y + \sigma I_m)\Phi^T$ such that

$$W\Phi(\Sigma_{Y|X'} + \sigma I_m)\Phi^T W^T = \mathrm{diag}\{\alpha_1,\ldots,\alpha_m\} \tag{46}$$

$$W\Phi(\Sigma_Y + \sigma I_m)\Phi^T W^T = I_m. \tag{47}$$

Claim that $\{\alpha_i\}_{i=1}^m$ are eigenvalues of the matrix $(\Sigma_Y + \sigma I_m)^{-\frac{1}{2}}(\Sigma_{Y|X'} + \sigma I_m)(\Sigma_Y + \sigma I_m)^{-\frac{1}{2}}$ and $U(\Phi) = (\Sigma_Y + \sigma I_m)^{\frac{1}{2}}\Phi^T W^T(\Phi)$ is formed by their corresponding eigenvectors.

To see this, it is straightforward to check that $U^T(\Phi)U(\Phi) = I_m$ and $U^T(\Phi)(\Sigma_Y + \sigma I_m)^{-\frac{1}{2}}(\Sigma_{Y|X'} + \sigma I_m)(\Sigma_Y + \sigma I_m)^{-\frac{1}{2}}U(\Phi) = \mathrm{diag}\{\alpha_1,\ldots,\alpha_m\}$.

We note that even though the eigenvalues of $(\Sigma_Y+\sigma I_m)^{-\frac{1}{2}}(\Sigma_{Y|X'}+\sigma I_m)(\Sigma_Y+\sigma I_m)^{-\frac{1}{2}}$ are independent of $\Phi$, the specific choice of the $m$ eigenvalues $\{\alpha_i\}_{i=1}^m$ does depend on $\Phi$. Hence $U(\Phi)$ is also a function of $\Phi$. Therefore, we end up with a fixed-point equation where the optimal orthogonal $\Phi^*$ must obey:

$$\Phi^* = W^{-T}(\Phi^*)U^T(\Phi^*)(\Sigma_Y+\sigma I_m)^{-\frac{1}{2}}. \tag{48}$$

The original ITM can be expressed as

$$\begin{aligned}&\arg\max_{\Phi} \mathcal{ITM}(\Phi,\beta),\\ &= \arg\min_{\Phi}\left\{\left(\frac{1}{\beta}-1\right)\log\,\det(\Phi\Sigma_Y\Phi^T+\Sigma)\right.\\ &\qquad\left.-\frac{1}{\beta}\log\,\det(\Phi\Sigma_{Y|X'}\Phi^T+\Sigma)\right\},\end{aligned} \tag{49}$$

$$\begin{aligned}&= \arg\min_{\Phi}\left\{\left(\frac{1}{\beta}-1\right)\log\,\det(W^{-1}W^{-T})\right.\\ &\qquad\left.-\frac{1}{\beta}\log\,\det(W^{-1}\mathrm{diag}\{\alpha_1,\ldots,\alpha_m\}W^{-T})\right\},\end{aligned} \tag{50}$$

$$\begin{aligned}&= \arg\min_{\Phi}\{-\log\,\det(W^{-1}W^{-T})\\ &\qquad-\frac{1}{\beta}\log\,\det(\mathrm{diag}\{\alpha_1,\ldots,\alpha_m\})\},\end{aligned} \tag{51}$$

$$= \arg\min_{\Phi}\left\{\log\,\det((W(\Phi))^2)-\frac{1}{\beta}\sum_{i\in\Gamma_1(\Phi)}\log\omega_i\right\}. \tag{52}$$

When $\beta\to 0^-$, the above ITM is essentially the solution of the following optimization:

$$\arg\max_{\Phi}\mathcal{ITM}(\Phi,\beta) = \arg\min_{\Phi}\left\{-\frac{1}{\beta}\sum_{i=1}^m\log\alpha_i\right\}, \tag{53}$$

$$= \arg\min_{\Phi}\left\{\sum_{i=1}^m\log\alpha_i\right\}, \tag{54}$$

where the minimum is achieved by choosing $\{\alpha_i\}_{i=1}^m$ as the smallest $m$ eigenvalues of $(\Sigma_Y+\sigma I_m)^{-\frac{1}{2}}(\Sigma_{Y|X'}+\sigma I_m)(\Sigma_Y+\sigma I_m)^{-\frac{1}{2}}$. Let $U_1$ be the matrix formed by the corresponding eigenvectors. The fixed-point equation (48) becomes

$$\Phi^* = W^{-T}(\Phi^*)U_1^T(\Sigma_Y+\sigma I_m)^{-\frac{1}{2}}. \tag{55}$$

Claim that the above fixed-point equation admits an analytical solution:

$$\Phi^* = \mathrm{orth}(U_1^T(\Sigma_Y+\sigma I_m)^{-\frac{1}{2}}). \tag{56}$$

It is enough to verify that $U^* = (\Sigma_Y+\sigma I_m)^{\frac{1}{2}}\Phi^{*T}W^{*T}(\Phi)$ is the eigenvector matrix corresponding to the smallest $m$ eigenvalues of $(\Sigma_Y+\sigma I_m)^{-\frac{1}{2}}(\Sigma_{Y|X'}+\sigma I_m)(\Sigma_Y+\sigma I_m)^{-\frac{1}{2}}$. By using the Gram-Schmidt method, we can rewrite $\Phi^* = LU_1^T(\Sigma_Y+\sigma I_m)^{-\frac{1}{2}}$, where $L\in\mathbb{R}^{m\times m}$ is a lower triangular matrix with positive diagonal entries. Hence, we have that $W(\Phi^*)=L^{-1}$ and the following expressions

$$\begin{aligned}&W(\Phi^*)\Phi^*(\Sigma_{Y|X'}+\sigma I_m)\Phi^{*T}W^T(\Phi^*),\\ &=\mathrm{diag}\{\alpha_1,\ldots,\alpha_m\}\end{aligned} \tag{57}$$

$$W(\Phi^*)\Phi^*(\Sigma_Y+\sigma I_m)\Phi^{*T}W^T(\Phi^*) = I_m. \tag{58}$$

Note that the optimization in (53) is independent of $W(\Phi)$. Therefore, we have that $\Phi^* = \mathrm{orth}(U_1^T(\Sigma_Y+\sigma I_m)^{-\frac{1}{2}})$

For the case when $\beta\to-\infty$, the ITM boils down to the following problem:

$$\arg\max_{\Phi}\mathcal{ITM}(\Phi,\beta) = \arg\min_{\Phi}\{-\log\,\det(\Phi\Sigma_Y\Phi^T+\Sigma)\}. \tag{59}$$

By the similar arguments, there exist a $W\in\mathbb{R}^{m\times m}$ such that

$$W(\Phi)W^T(\Phi) = \mathrm{diag}\{\alpha_1,\ldots,\alpha_m\}, \tag{60}$$

$$W(\Phi)\Phi(\Sigma_Y+\sigma I_m)\Phi^TW^T(\Phi) = I_m. \tag{61}$$

Further, $\{\alpha_i\}_{i=1}^m$ are eigenvalues of the matrix $(\Sigma_Y+\sigma I_m)^{-1}$ and $U(\Phi)=(\Sigma_Y+\sigma I_m)^{\frac{1}{2}}\Phi^TW^T(\Phi)$ is formed by their corresponding eigenvectors. Hence, we have that

$$\begin{aligned}&\arg\max_{\Phi}\mathcal{ITM}(\Phi,\beta)\\ &=\arg\min_{\Phi}\{-\log\,\det(\Phi\Sigma_Y\Phi^T+\Sigma)\},\end{aligned} \tag{62}$$

$$= \arg\min_{\Phi}\{-\log\,\det(W^{-1}W^{-T})\}, \tag{63}$$

$$= \arg\min_{\Phi}\{\sum_{i=1}^m\log\alpha_i\}. \tag{64}$$

In order to solve the above optimization, we require that $\{\alpha_i\}_{i=1}^m$ are the smallest $m$ eigenvalues of $(\Sigma_Y+\sigma I_m)^{-1}$ and the optimal projection matrix is given by

$$\Phi^* = \mathrm{orth}(U_2^T(\Sigma_Y+\sigma I_m)^{-\frac{1}{2}}), \tag{65}$$

where $U_2$ is constituted by the eigenvectors corresponding to the smallest $m$ eigenvalues of the matrix $(\Sigma_Y+\sigma I_m)^{-1}$. □

## REFERENCES

[1] M. Seeger and H. Nickisch, "Compressed sensing and Bayesian experimental design," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 912–919.

[2] K. Torkkola, "Learning discriminative feature transforms to low dimensions in low dimensions," in *Proc. Int. Conf. Adv. Neural Inform. Process. Syst.*, 2001, pp. 969–976.

[3] K. Torkkola, "Feature extraction by non parametric mutual information maximization," *J. Mach. Learning Res.*, vol. 3, pp. 1415–1438, 2003.

[4] M. Chen, W. Carson, M. Rodrigues, R. Calderbank, and L. Carin, "Communications inspired linear discriminant analysis," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 919–926.
[5] L. Song, A. Smola, K. Borgwardt, and A. Gretton, "Colored maximum variance unfolding," in *Proc. Int. Conf. Adv. Neural Inform. Process. Syst.*, 2008, pp. 1385–1392.
[6] E. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.
[7] M. Gehm, R. John, D. Brady, R. Willett, and T. Schulz, "Single-shot compressive spectral imaging with a dual-disperser architecture," *Optics Exp.*, vol. 15, no. 21, pp. 14 013–14 027, 2007.
[8] W. Carson, M. Chen, M. Rodrigues, R. Calderbank, and L. Carin, "Communications-inspired projection design with application to compressive sensing," *SIAM J. Imag. Sci.*, vol. 5, no. 4, pp. 1185–1212, 2012.
[9] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, Jun. 2008.
[10] K. Hild, D. Erdogmus, K. Torkkola, and J. Principe, "Feature extraction using information-theoretic learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1385–1392, Sep. 2006.
[11] S. Kaski and J. Peltonen, "Informative discriminant analysis," in *Proc. 25th Int. Conf. Mach. Learn.*, 2003, pp. 329–336.
[12] Z. Nenadic, "Information discriminant analysis: Feature extraction with an information-theoretic objective," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1394–1407, Aug. 2007.
[13] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun., Control, Comput.*, 1999, pp. 368–377.
[14] N. Tishby and N. Slonim, "Data clustering by Markovian relaxation and the information bottleneck method," in *Proc. Int. Conf. Adv. Neural Inform. Process. Syst.*, 2000, pp. 640–646.
[15] N. Slonim, N. Friedman, and N. Tishby, "Multivariate information bottleneck," *Neural Comput.*, vol. 18, no. 8, pp. 1739–1789, 2006.
[16] W. Hsu, L. Kennedy, and S. F. Chang, "Video search reranking via information bottleneck principle," in *Proc. ACM Int. Conf. Multimedia*, 2006, pp. 35–44.
[17] F. Creutzig and H. Sprekeler, "Predictive coding and the slowness principle: An information-theoretic approach," *Neural Comput.*, vol. 20, no. 4, pp. 1026–1041, 2008.
[18] R. Hecht, E. Noor, and N. Tishby, "Speaker recognition by Gaussian information bottleneck." in *Proc. INTERSPEECH*, 2009, pp. 1567–1570.
[19] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for Gaussian variables," *J. Mach. Learn. Res.*, vol. 6, no. 1, pp. 165–188, 2005.
[20] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.
[21] L. Wang, M. Rodrigues, and L. Carin, "Generalized Bregman divergence and gradient of mutual information in vector Poisson channels," in *Proc. IEEE Int. Symp. Inform. Theory Proc.*, 2013, pp. 454–458.
[22] L. Wang, D. Carlson, M. Rodrigues, D. Wilcox, R. Calderbank, and L. Carin, "Designed measurements for vector count data," in *Proc. Int. Conf. Adv. Neural Inform. Process. Syst.*, 2013, pp. 1142–1150.
[23] L. Wang, D. Carlson, M. Rodrigues, R. Calderbank, and L. Carin, "A Bregman matrix and the gradient of mutual information for vector Poisson and Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 60, no. 5, pp. 2611–2629, May 2014.
[24] B. Tang, J. Tang, and Y. Peng, "MIMO radar waveform design in colored noise based on information theory," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4684–4697, Sep. 2010.
[25] L. Zheng and D. Tse, "Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. Inform. Theory*, vol. 48, no. 2, pp. 359–383, Feb. 2002.
[26] R. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
[27] D. Wilcox, G. Buzzard, B. Lucier, P. Wang, and D. Ben-Amotz, "Photon level chemical classification using digital compressive detection," *Analytica Chimica Acta*, vol. 755, pp. 17–27, 2012.
[28] N. Halko, P. Martinsson, and J. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Rev.*, vol. 53, no. 2, pp. 217–288, 2011.
[29] C. Xiao, Y. Zheng, and Z. Ding, "Globally optimal linear precoders for finite alphabet signals over complex vector Gaussian channels," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3301–3314, Jul. 2011.
[30] S. Prasad, Certain relations between mutual information and fidelity of statistical estimation (2012) [Online]. *Available: http://arxiv.org/pdf/1010.1508v1.pdf*
[31] M. Hellman and J. Raviv, "Probability of error, equivocation, and the Chernoff bound," *IEEE Trans. Inform. Theory*, vol. 16, no. 4, pp. 368–372, Jul. 1970.
[32] D. Palomar and S. Verdú, "Representation of mutual information via input estimates," *IEEE Trans. Inform. Theory*, vol. 53, no. 2, pp. 453–470, Feb. 2007.
[33] C. Bishop and N. Nasrabadi, *Pattern Recognition and Machine Learning*. vol. 1, Berlin, Germany: Springer, 2006.
[34] A. Lozano, A. Tulino, and S. Verdú, "Optimum power allocation for parallel Gaussian channels with arbitrary input distributions," *IEEE Trans. Inform. Theory*, vol. 52, no. 7, pp. 3033–3051, Jul. 2006.
[35] F. Pérez-Cruz, M. Rodrigues, and S. Verdú, "MIMO Gaussian channels with arbitrary inputs: Optimal precoding and power allocation," *IEEE Trans. Inform. Theory*, vol. 56, no. 3, pp. 1070–1084, Mar. 2010.
[36] 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," *IEEE Trans Signal Process.*, vol. 58, no. 12, pp. 6140–6155, Dec. 2010.
[37] M. Zhou, L. Hannah, D. Dunson, and L. Carin, "Beta-negative binomial process and Poisson factor analysis," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2012, pp. 1462–1471.
[38] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[39] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.
[40] A. Gut, *An Intermediate Course Probability*. vol. 1, Berlin, Germany: Springer, 1995.

**Liming Wang** (S08-M11) received the BS degree in electronic engineering from the Huazhong University of Science and Technology, China, in 2006, MS degree in mathematics and the PhD degree in electrical and computer engineering from the University of Illinois at Chicago, both in 2011. From 2011 to 2012, he was a postdoctoral research scientist in the Department of Electrical Engineering, Columbia University. Since May 2012, he has been a postdoctoral associate in the Department of Electrical and Computer Engineering, Duke University. His research interests are in high-dimensional data processing; machine learning; compressive sensing; statistical signal processing; bioinformatics and genomic signal processing.

**Minhua Chen** received the BS degree in electronic engineering from Tsinghua University in 2004, and the MS and PhD degrees in electrical and computer engineering from Duke University in 2009 and 2012. He was a postdoctoral researcher at the University of Chicago from September 2012 to May 2014. He is currently a research scientist at Amazon. His research interest is in machine learning and signal processing.

**Miguel R. D. Rodrigues** (S98-A02-M03) received the Licenciatura degree in electrical engineering from the University of Porto, Portugal in 1998 and the PhD degree in electronic and electrical engineering from the University College London, United Kingdom, in 2002. He is currently a senior lecturer with the Department of Electronic and Electrical Engineering, University College London, United Kingdom. He was previously with the Department of Computer Science, University of Porto, Portugal, rising through the ranks from assistant to associate professor, where he also led the Information Theory and Communications Research Group, Instituto de Telecomunicaes - Porto. He has held postdoctoral or visiting appointments with various Institutions worldwide including University College London, Cambridge University, Princeton University, and Duke University in the period 2003-2013. His research work, which lies in the general areas of information theory, communications and signal processing, has led to more than 100 papers in journals and conferences to date. He was also honored by the IEEE Communications and Information Theory Societies Joint Paper Award in 2011 for his work on Wireless Information-Theoretic Security (joint with M. Bloch, J. Barros and S. M. McLaughlin).

**David Wilcox** received the PhD degree in chemistry from Purdue University in 2012. He is currently a postdoctoral researcher in chemistry at the University of Chicago. His research focus is on optical chemical sensing and compressive sensing.

**Robert Calderbank** (M89-SM97-F98) received the BSc degree in 1975 from the Warwick University, England, the MSc degree in 1976 from the Oxford University, England, and the PhD degree in 1980 from the California Institute of Technology, all in mathematics. He is currently a professor of electrical engineering at Duke University where he now directs the Information Initiative at Duke (iiD) after serving as dean of Natural Sciences (2010-2013). He was previously professor of Electrical Engineering and Mathematics, Princeton University where he directed the Program in Applied and Computational Mathematics. Prior to joining Princeton in 2004, he was vice president for Research at AT&T, responsible for directing the first industrial research lab in the world where the primary focus is data at scale. At the start of his career at Bell Labs, innovations by him were incorporated in a progression of voiceband modem standards that moved communications practice close to the Shannon limit. Together with Peter Shor and colleagues at AT&T Labs he showed that good quantum error correcting codes exist and developed the group theoretic framework for quantum error correction. He is a co- inventor of space-time codes for wireless communication, where correlation of signals across different transmit antennas is the key to reliable transmission. He served as editor in chief of the *IEEE Transactions Information Theory* from 1995 to 1998, and as associate editor for Coding Techniques from 1986 to 1989. He was a member of the Board of Governors of the IEEE Information Theory Society from 1991 to 1996 and from 2006 to 2008. He received the IEEE Information Theory Prize Paper Award in 1995 for his work on the Z4 linearity of Kerdock and Preparata Codes (joint with A. R. Hammons Jr., P. V. Kumar, N. J. A. Sloane, and P. Sole), and again in 1999 for the invention of space-time codes (joint with V.Tarokh and N. Seshadri). He has received the 2006 IEEE Donald G. Fink Prize Paper Award, the IEEE Millennium Medal, the 2013 IEEE Richard W. Hamming Medal, and he was elected to the US National Academy of Engineering in 2005.

**Lawrence Carin** (SM'96-F'01) received the BS, MS, and PhD degrees in electrical engineering from the University of Maryland, College Park, in 1985, 1986, and 1989, respectively. In 1989 he joined the Electrical Engineering Department, Polytechnic University (Brooklyn) as an assistant professor, and became an associate professor there in 1994. In September 1995, he joined the Electrical Engineering Department, Duke University, where he is currently a professor. He was chairman of the Electrical and Computer Engineering Department from 2011-2014. He held the William H. Younger Professorship from 2003-2013. He is currently Vice Provost for Research at Duke. He is a co-founder of Signal Innovations Group, Inc. (SIG), a small business that was acquired by BAE Systems in 2014. His current research interests include machine learning and applied statistics. He has published more than 300 peer-reviewed papers, and he is a member of the Tau Beta Pi and Eta Kappa Nu honor societies.