Nonlinear Information-Theoretic Compressive Measurement Design

Liming Wang[1] LIMING.W@DUKE.EDU
Abolfazl Razi[1] ABOLFAZL.RAZI@DUKE.EDU
Miguel Dias Rodrigues[2] M.RODRIGUES@UCL.AC.UK
Robert Calderbank[1] ROBERT.CALDERBANK@DUKE.EDU
Lawrence Carin[1] LCARIN@DUKE.EDU

[1]Department of Electrical & Computer Engineering, Duke University, Durham, NC 27708, USA
[2]Department of Electronic and Electrical Engineering, University College London, London, UK

Abstract

We investigate design of general nonlinear functions for mapping high-dimensional data into a lower-dimensional (compressive) space. The nonlinear measurements are assumed contaminated by additive Gaussian noise. Depending on the application, we are either interested in recovering the high-dimensional data from the nonlinear compressive measurements, or performing classification directly based on these measurements. The latter case corresponds to classification based on nonlinearly constituted and noisy features. The nonlinear measurement functions are designed based on constrained mutual-information optimization. New analytic results are developed for the gradient of mutual information in this setting, for arbitrary input-signal statistics. We make connections to kernel-based methods, such as the support vector machine. Encouraging results are presented on multiple datasets, for both signal recovery and classification. The nonlinear approach is shown to be particularly valuable in high-noise scenarios.

1. Introduction

Dimensionality reduction plays a pivotal role in numerous machine-learning applications, including compressive measurements and feature design (Seeger & Nickisch, 2008; Chen et al., 2012; Wang et al., 2013; 2014). Among those approaches, linear dimensionality reduction has gained popularity, due to its relatively simple formulation and theoretical analysis (Candès et al., 2006). Linear measurements may be analyzed in terms of multiplying a signal of interest with a measurement matrix. Ideally the number of rows of this matrix is small relative to the dimension of the original data vector, and Gaussian additive measurement noise is often assumed (Carson et al., 2012; Ji et al., 2008). However, the restriction to linear measurements is limiting, in that many measurement systems are inherently nonlinear. Further, in the context of feature design, the assumption of linear features may limit discrimination quality. The counterpart to linear measurements is to replace each row of the aforementioned measurement matrix by an associated nonlinear measurement function (Jarrett et al., 2009; Karklin & Simoncelli, 2011; Xu et al., 2013). Rather than designing multiple rows of a linear measurement matrix, the objective is to design a set of nonlinear measurement functions. Nonlinear dimensionality reduction techniques have exhibited better performances and flexibility (Schölkopf et al., 1998; Tenenbaum et al., 2000; Song et al., 2008), relative to linear measurements. However, there has been far less work in the literature on designing these multiple nonlinear measurement functions, relative to the vast literature on linear measurements.

Numerous existing nonlinear dimensionality-reduction methods (Schölkopf et al., 1998; Song et al., 2008) are essentially manifested via the *kernel trick* or *kernel method* (Aizerman et al., 1964), whose idea can be briefly summarized as follows. We consider the classification case, but similar issues hold for regression. The original data is believed to *not* be linearly separable, $i.e.$, there is not a hyperplane that separates the classes of data. Hence, a nonlinear function is desired, that maps the original data to a new feature space of a higher dimension (possibly *much* higher, even of infinite dimension). The transformed data ideally becomes linearly separable in the new high-dimensional space, in which methods such as Principal Component Analysis (PCA) or linear Support Vector Machine (SVM) can be readily employed (Shawe-Taylor & Cristianini, 2004) (linear methods, applied *after* nonlinear transformation). These methods rely on inner products in the high-dimensional space, and these inner products are

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

replaced by a Mercer kernel (Aizerman et al., 1964). This is the so-called "kernel trick," in that via invoking a Mercer kernel to represent inner products, one never has to explicitly design or implement the nonlinear mapping function.

The kernel trick significantly simplifies the computations involved in the nonlinear map and may provide a substantial performance improvement. However, such an improvement is achieved only when a suitable kernel function is selected, thereby requiring a sophisticated kernel design method to assure good performance. In cases for which a special nonlinear structure is needed (Karklin & Simoncelli, 2011), structural constraints are often difficult to explicitly impose in the kernel trick.

Another approach for nonlinear dimensionality reduction is to directly model or learn the nonlinear compressive function, and design it in a way that the compressed data maximally conveys a desired form of information. Among various information-theoretic metrics, mutual information is widely utilized (Hild et al., 2006; Kaski & Peltonen, 2003; Nenadic, 2007). Mutual-information-based *linear* projection design for the Gaussian measurement model has been considered in (Carson et al., 2012) for signal recovery, and in (Chen et al., 2012) for classification (feature design). Similar linear projection design has been considered for the Poisson model (Wang et al., 2013).

We present new theoretical results for gradient of mutual information under the *nonlinear* measurement model, for both signal recovery and classification, extending previous results that assumed linear measurements (Guo et al., 2008; Carson et al., 2012; Chen et al., 2012). The results for the assumption of a linear measurement are recovered as a special case of our results. Our theoretical results assume an *arbitrary* distribution for the source, and can be applied to a broad range of applications. In addition to the theoretical contributions, we demonstrate how the results may be used in practice, by providing numerical results in the context of compressive image sensing and multi-class classification. We demonstrate on multiple datasets that designed nonlinear measurement functions can yield improved performance relative to linear and random projections as well as the kernel SVM method, especially in the relatively low signal-to-noise ratio (SNR) regime.

2. Nonlinear Measurement Model

2.1. Problem Statement

Assume the data $X \in \mathbb{R}^n$ is drawn from the distribution P_X. In the case for which there is an underlying class label, $P_X = \sum_{i=1}^T P_C(C = i)P_{X|C}(X|C = i) = \sum_{i=1}^T \pi_i P_{X|C}(X|C = i)$, where C is the class label, T is the total number of classes and $C \sim \sum_{i=1}^T \pi_i \delta_i$. We do not assume a specific form of $P_{X|C}$ and thus a general mixture model for X is considered. We do assume that P_C and $P_{X|C}$ are known or can be estimated from training data.

The nonlinear measurement Y is modeled as

$$Y = \Phi(X) + W, \tag{1}$$

where $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ is a nonlinear measurement function with (ideally) $m \ll n$ and $W \sim \mathcal{N}(\mathbf{0}, \Sigma)$ is additive Gaussian noise with zero mean and covariance matrix Σ. We further assume that Φ admits an expansion under some basis. In this paper, we mainly focus on the polynomial basis, $i.e.$, we consider the Taylor expansion of Φ. However, our theoretical result is valid for arbitrary basis.

Consider a polynomial expansion up to kth order:

$$\Phi(X) = \begin{bmatrix} \Phi_1(X) \\ \vdots \\ \Phi_m(X) \end{bmatrix} = \begin{bmatrix} \sum_{1\le|\alpha|\le k} a_\alpha^{(1)} X^\alpha \\ \vdots \\ \sum_{1\le|\alpha|\le k} a_\alpha^{(m)} X^\alpha \end{bmatrix}, \tag{2}$$

where $\alpha = [\alpha_1, \ldots, \alpha_n] \in \mathbb{Z}_+^n$ is the multi-index; $X^\alpha := X_1^{\alpha_1} \times \cdots \times X_n^{\alpha_n}$ and $|\alpha| := \sum_i \alpha_i$. If we set $k = 1$, this reduces to a linear measurement model, like that considered widely previously (Seeger & Nickisch, 2008; Chen et al., 2012; Wang et al., 2013; 2014; Candès et al., 2006; Carson et al., 2012).

Let $A \in \mathbb{R}^{m\times r}$ denote the coefficient matrix, where the i-th row of A is consecutively constituted by $a_\alpha^{(i)}$ with the dictionary ordering on α, $i.e.$, $A_i = [a^{(i)}_{[1,0,\ldots,0]}, a^{(i)}_{[0,1,\ldots,0]}, \ldots, a^{(i)}_{[0,0,\ldots,k]}]$. Similarly, we denote $\psi(X)$ as the polynomial basis vector where each entry is sequentially composed by X^α with the dictionary ordering, $i.e.$, $\psi(X) = [X_1, X_2, \ldots, X_n, X_1^2, X_1X_2, \ldots, X_n^k]^T$; there are r terms in $\psi(X)$. In this manner, we can rewrite the nonlinear measurement function Φ as

$$\Phi(X) = A\psi(X). \tag{3}$$

Note that $\psi(X)$ is fixed once we set the basis and the expansion order (while above and in our experiments we focus on polynomial expansions, the same ideas hold for arbitrary expansions by which $\psi(X)$ is constituted).

We wish to design Φ with the goal of maximizing the mutual information between random variables X and Y. This is termed the "signal recovery" problem, as our goal is to recover the input signal X. Alternatively, we may be interested in problems for which $X|C \sim P_{X|C}$ with $C \sim \sum_{i=1}^T \pi_i\delta_i$. Now the objective is to design Φ to maximize the information content in Y about the class label C; we term this the "classification" problem.

In order to design the nonlinear $\Phi^\star$ for signal recovering, with $X \sim P_X$ and $Y|X \sim \mathcal{N}(\Phi(X), \Sigma)$, we adopt the criterion

$$\Phi^\star = \arg\max_{\Phi} I(X;Y) \tag{4}$$
$$= \arg\max_{A} I(X;Y). \tag{5}$$

where $I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$ is the mutual information between X and Y, $h(X)$ is the differential entropy, and $h(X|Y)$ is the conditional differential entropy (Cover & Thomas, 2006).

For the classification problem, we adopt the criterion

$$\Phi^\star = \arg\max_{\Phi} I(C;Y) \tag{6}$$
$$= \arg\max_{A} I(C;Y). \tag{7}$$

where $I(C;Y) = H(C) - H(C|Y)$, where $H(C)$ and $H(C|Y)$ are the entropy and conditional entropy, respectively (Cover & Thomas, 2006).

We consider the above optimization problems in the context of the energy constraint

$$\mathbb{E}\{\text{tr}[\Phi\Phi^T]\} \leq E. \tag{8}$$

with E being a constant. The Shannon entropy (and Shannon differential entropy) with natural logarithm is considered throughout (Cover & Thomas, 2006).

The mutual-information-based criterion for signal recovery may be justified by noting that it has been shown recently that (Prasad, 2012)

$$\text{MMSE} \geq \frac{1}{2\pi e}\exp\{2[h(Y) - I(X;Y)]\} \tag{9}$$

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

The mutual information metric for classification is justified by recalling the Bayesian classification error $P_e = \int P_Y(y)[1 - \max_c P_{C|Y}(c|y)]dy$, and noting that it has been shown in (Hellman & Raviv, 1970) that

$$P_e \leq \frac{1}{2}H(C|Y) \tag{10}$$

where $H(C|Y) = H(C) - I(C;Y)$. Since $H(C)$ is independent of Φ, minimizing the upper bound to P_e is equivalent to maximizing $I(C;Y)$.

3. Gradients of Mutual Information

3.1. Explicit Gradient Formulas

The mutual information terms in (5) and (7) generally do not possess known analytic form for general source statistics. Rather than evaluating the mutual information directly, we consider the gradient of mutual information with respect to A. This sheds light on solving this class of optimization problems, and it may be used in numerical experiments (gradient-based design). It is desirable to derive an analytical form of this gradient, if possible. In this section, we present two theorems on the gradient of mutual information for the nonlinear measurement model. We always assume the regularity conditions, specifically, that the order of integration and differentiation can be interchanged freely, and the expectation operator $\mathbb{E}(\cdot)$ may be interchanged. This assumption is mild and almost always valid in practice (Palomar & Verdú, 2007; Wang et al., 2014).

Theorem 1. *Assuming the regularity conditions, the gradient of mutual information $I(X;Y)$ with respect to A, for the nonlinear measurement model in (1), can be expressed as*

$$\nabla_A I(X;Y) = \tag{11}$$
$$\Sigma^{-1}A\mathbb{E}[[\psi(X) - \mathbb{E}[\psi(X)|Y]][\psi(X) - \mathbb{E}[\psi(X)|Y]]^T].$$

The above theorem generalizes the scalar result for nonlinear measurement model in (Guo et al., 2005a), and the linear case (Guo et al., 2005b; Carson et al., 2012) now becomes a corollary of Theorem 1 where we set $\psi(X) = X = [X_1, \ldots, X_n]^T$ and $A \in \mathbb{R}^{m\times n}$.

Corollary 1. *Assuming the regularity conditions, the gradient of mutual information $I(X;Y)$ for the linear measurement model $Y = AX + W$, where $W \sim \mathcal{N}(\mathbf{0}, \Sigma)$ can be expressed as*

$$\nabla_A I(X;Y) = \tag{12}$$
$$\Sigma^{-1}A\mathbb{E}[[X - \mathbb{E}[X|Y]][X - \mathbb{E}[X|Y]]^T].$$

The gradient of mutual information between the class label and the measurement can also be established as follow.

Theorem 2. *Assuming the regularity conditions, the gradient of mutual information $I(C;Y)$ for the nonlinear measurement model in (1) can be expressed as*

$$\begin{aligned}\nabla_A I(C;Y) &= \Sigma^{-1}A \\ &\times \mathbb{E}\{[\mathbb{E}[\psi(X)|Y,C] - \mathbb{E}[\psi(X)|Y]] \\ &\times [\mathbb{E}[\psi(X)|Y,C] - \mathbb{E}[\psi(X)|Y]]^T\}.\end{aligned} \tag{13}$$

A corollary to the above theorem, for the linear measurement special case, yields the results in (Chen et al., 2012). The proofs of the above theorems are presented in the Supplementary Material. Note that Theorems 1 and 2 are valid for *arbitrary* P_X or $P_{X|C}$, as well as *arbitrary* non-linear basis function ψ. We also emphasize here that albeit its resemblance to the linear case, Theorems 1 and 2 can not be easily deduced from their linear counterparts, especially when ψ is not injective.

3.2. Discussion of the Theorems

The above theorems indicate that the gradient of mutual information for nonlinear mappings is closely related to the MMSE matrix of the *transformed* signal $\psi(X)$ which, for $k > 1$ in (2), resides in a much higher-dimensional space than the original X. Specifically, note the MMSE matrices $\mathbb{E}[[\psi(X) - \mathbb{E}[\psi(X)|Y]][\psi(X) - \mathbb{E}[\psi(X)|Y]]^T]$ for signal recovery, and $\mathbb{E}\{[\mathbb{E}[\psi(X)|Y, C] - \mathbb{E}[\psi(X)|Y]][\mathbb{E}[\psi(X)|Y, C] - \mathbb{E}[\psi(X)|Y]]^T\}$ for classification (for the classification case, this is a *generalized* MMSE matrix, as in (Chen et al., 2012)). The MMSE matrices characterize optimal estimation accuracy in the high-dimensional space defined by $\psi(X)$, and the explicit relationships for the gradients of mutual information interrelate information-theoretic and estimation-based metrics, generalizing (Guo et al., 2008).

We have m nonlinear measurements, defined by $(\Phi_1(X), \dots, \Phi_m(X))$, ideally with $m \ll n$, where $X \in \mathbb{R}^n$. Although when $m < n$ we manifest a compressive measurement (in that the dimension of Y is smaller than the dimension of X), the design procedure of that measurement may be viewed as first performing a high-dimensional mapping. Specifically, the map $X \to \psi(X)$ translates X to an (often much) higher-dimensional space, and this space is shared for all $(\Phi_1(X), \dots, \Phi_m(X))$. The aforementioned MMSE matrices characterize optimal estimation in that higher-dimensional space, characterized by the mapping of the statistics of X to the statistics of $\psi(X)$. While traditional compressive sensing is characterized by a linear mapping directly from X to Y, we here first manifest a nonlinear mapping to a high-dimensional space, via $\psi(X)$. Once in that higher dimensional space, the subsequent measurement is like in traditional linear compressive sensing (Candès et al., 2006).

Because once the mapping $\psi(X)$ is performed everything proceeds like in traditional compressive sensing, Theorem 1 looks like the results in (Carson et al., 2012), and Theorem 2 looks like the results in (Chen et al., 2012); the only difference appearing to be that here $\psi(X)$ replaces X in the previous work. Similar connections appear with respect to the results in (Guo et al., 2008). However, we emphasize that the proof of the theorems is *not* a direct application of these previous results. To see this, note that $X \to \psi(X) \to Y$ forms a Markov chain, and via the Data Processing Inequality (Cover & Thomas, 2006), we have that $I(X; Y) \leq I(\psi(X), Y)$. The equality may not generally hold when ψ is not injective, thereby invalidating an easy argument following directly via the gradient of linear model $\nabla_A I(\psi(X); Y)$.

Instead, by assuming a general ψ, our proofs are done by computing directly the gradient of the mutual information associated with the measurement model in (1), (2) and (3), with details presented in the Supplementary Material. Beyond the fact that this theorem represents a generalization of previous results associated with intersections between information theory and estimation theory to nonlinear settings, thereby filling an important gap in the literature, the significance of the theorem is also twofold. First, it describes how changes in the parameters of the nonlinear expansions affect the mutual information (associated with recovery or classification problems) via generalisations of the MMSE matrix (for recovery) or the equivalent MMSE matrix (for classification) in the higher-dimensional space induced by the expansions: this then provides the means to articulate in an information-theoretic manner about the value of kernel methods, as discussed later. Second, the theorem also provides a mechanism to perform nonlinear measurement designs in compressive settings: in fact, the practical value of the theorem is justified by state-of-the-art results in various recovery and classification problems, as shown in Section 5.

3.3. Gradient-Based Numerical Design

A numerical solution to the optimization problems in (5) and (7) can be realized via a gradient-descent method. The MMSE matrices involved in Theorems 1 and 2 can be readily calculated by Monte Carlo integration, as in (Carson et al., 2012; Chen et al., 2012); we elaborate on this calculation when presenting experimental results. We summarize the algorithm as follows:

1. Select suitable basis vector $\psi(X)$ and initialize A.
2. Use Monte Carlo integration to calculate the MMSE matrices involved in Theorems 1 and 2. Update the A matrix as $A^{new} = \text{proj}(A^{old} + \delta \nabla_A I(\cdot, Y)$, where δ is the step size, $I(\cdot, Y)$ is the mutual information of interest and proj$(\cdot)$ projects the matrix to the feasible set defined by the energy constraint in (8), *i.e.*, renormalize A to satisfy the energy constraint.
3. Repeat previous step until convergence.

In general the mutual information is not a concave function of A, and therefore we cannot guarantee a global-optimal solution. In all experiments the solution converged to a useful/effective solution from a random start.

4. Relationship to Previous Work

4.1. Connection to Kernel Methods

We have noted in Section 3.2 that the proposed procedure may be viewed as the sequence $X \to \psi(X) \to A\psi(X)$, where we first map X to a higher-dimensional space via $\psi(X)$, and then perform a linear measurement via A to the low-dimensional space in which the m-dimensional measurement is performed. Methods like the SVM are *motivated* by first implementing a similar mapping $\psi(X)$, but the mapping is not applied explicitly, as inner products in the high-dimensional space are replaced by a Mercer kernel.

However, some kernel methods, such as kernel PCA

(Schölkopf et al., 1998), can be regarded as a special case of the proposed approach. To see this, let $\phi(x) : \mathbb{R}^n \to \mathbb{R}^r$ with $r \gg n$ be the implicit feature mapping used in the kernel methods. By the dense property of the polynomial functions (Folland, 1999), we have that $\phi(x) \approx A_\phi \psi(x)$, where A_ϕ is the associated coefficient matrix, $\psi(x)$ is the polynomial basis vector defined in the previous section, and the accuracy of this approximation is improved with the increase of the polynomial order. Afterwards, a linear projection matrix P is applied to obtain the compressed measurement $\Phi(x) = PA_\phi\psi(x)$. It is straightforward to see that by equating $A = PA_\phi$, the kernel methods fall into a special case of our approach. Compared to the kernel methods, in which one has to implicitly calculate the coefficient matrix A_ϕ via a Mercer's kernel function and constitute the projection matrix P indirectly via the kernel trick, by relaxing PA_ϕ to one matrix A, it only requires one to specify a basis ψ in which a linear separation of the data is *possible*. Furthermore, such a requirement is almost always guaranteed with a high enough expansion order (high enough order of the Taylor expansion under the polynomial basis).

4.2. Connection to Compressive Sensing

As discussed in Section 3.2, the proposed procedure may be viewed as a mapping $X \to \psi(X)$, followed by linear measurements to manifest $\psi(X) \to Y$, with the linear measurements (projections) defined by the rows of A. This is a generalization of compressive sensing (CS) (Candès & Wakin, 2008; Candès et al., 2006), in that in CS-like linear measurements are also performed. However, here the linear measurements are performed *after* manifesting a mapping to a higher-dimensional space, via $\psi(X)$. Another distinction with CS is that in the original theory the projection matrix A was constituted at random, however here we design A. It was demonstrated in (Carson et al., 2012; Chen et al., 2012) that such designed CS measurements often yield better results than randomly constituted A, and he we extend this concept to linear measurements after a mapping to a higher-dimensional space.

An important new theoretical result was developed in (Blumensath, 2013), in which the same nonlinear measurement model $Y = \Phi(X) + W$ is considered, where $\Phi(X)$ is a set of m nonlinear functions, like we have considered. The author linearizes $\Phi(X)$ by taking the first order expansion (like our first derivative), and this expansion yields a first-order measurement model AX, where A is a function of X. Recovery guarantees are derived assuming X is sparse, and under the condition that A satisfies restricted isometry property (RIP) conditions. Here we design the $\Phi(X)$ for all X characterized by distribution P_X (we do not locally linearize), and we consider general P_X (do not require sparsity). Further, we consider this design for the classification problem as well, not addressed in (Blumensath, 2013). While we do not have performance guarantees, (9) and (10) provide theoretical assurances on the quality of the subsequent results, validated in our experiments.

5. Experiments

We evaluate the proposed nonlinear measurement design for both signal recovery and classification applications. The gradient results in Theorems 1 and 2 are valid for arbitrary mixture distribution of X and can be readily implemented via Monte Carlo integration, provided the posterior $P_{\psi(X)|Y}$ can be easily sampled (we use a Gaussian mixture model (GMM) signal model to achieve that goal). We assume $P_X(x) = \sum_{i=1}^T \pi_i P_{X|C=i}$ for input data X and employ a GMM for $P_{\psi(X)|C=i}(x) = \sum_{j=1}^{N_i} \pi_i^{(j)} \mathcal{N}(x; \mu_i^{(j)}, \Sigma_i^{(j)})$. Consequently, the distribution of $\psi(X)$ for each discrete class label $C = i$ is a GMM. The distribution $P_{\psi(X)}$ is also a GMM with the class label C summed out: $P_{\psi(X)}(x) = \sum_{i=1}^T \sum_{j=1}^{N_i} \pi_i \pi_i^{(j)} \mathcal{N}(x; \mu_i^{(j)}, \Sigma_i^{(j)})$, where $\pi_i^{(j)}$, $j = 1, \dots, N_i$ are the GMM coefficients within class i. $\mu_i^{(j)}$ and $\Sigma_i^{(j)}$, $j = 1, \dots, N_i$ are the means and variances of respective N_i Gaussian components. The GMMs are learned as discussed in (Chen et al., 2010), based on training data $\{\psi(X_i)\}$, obtained simply by apply the function $\psi(X)$ on the input training data $\{X_i\}$. The nonparametric method in (Chen et al., 2010) used to learn the GMMs infers the number of needed mixture components, via use of the Dirichlet process. As is well known, the GMM is able to capture the true underlying distribution of $\psi(X)$ with increased number of components N_i, and therefore the assumption of a GMM source is not limiting in practice.

In addition to being easily sampled, the GMM representation has the advantage of an *analytic* posterior $P_{\psi(X)|Y}$, which is also a GMM, as detailed in (Chen et al., 2010; 2012). In particular, we have that $P_{\psi(X)|Y} = \sum_{i=1}^T \tilde{\pi}_i P_{\psi(X)|Y,C=i}$, with analytic expressions for $\{\tilde{\pi}_i\}$ and $P_{\psi(X)|Y,C=i}$. Within the Taylor/polynomial expansion $\psi(X)$, the first-order terms in $\psi(X)$ are X itself. Therefore, when performing estimation of X based on observed Y, we first (analytically) compute $P_{\psi(X)|Y}$, and we then marginalize out all components other than those associated with X itself. The MMSE estimated recovered signal is $\mathbb{E}[X|Y]$, computed efficiently in closed form. Since the posterior $P_{X|Y}$ is also a GMM, it naturally induces a Bayesian classifier $\max_c P_{C=c|Y}$, where $P_{C=c|Y} = \tilde{\pi}_c$.

To obtain faster convergence in the gradient descent method using Theorems 1 and 2, we normalize the basis vector $\psi(X)$ by apply a constant vector $\gamma \in \mathbb{R}^r$ such that the $\mathbb{E}[\text{diag}\big((\gamma \circ \psi(X)(\gamma \circ \psi(X))^T\big)] = \mathbf{1}$, where $\circ$ denotes the Hadamard product (entry-wise product), $\mathbf{1}$ is a $r \times 1$ all-one vector and r is the dimension of $\psi(X)$, as previously defined. This normalization balances the contributing nonlinear terms of different degrees, and avoids the appearance of any dominating terms. The normalizing vector γ can be empirically obtained via the sample covariance of training data $\{\psi(X)\}$. Moreover, to avoid extra computational cost brought by the very high dimensional vector $\psi(X)$, for large k, we reduce the dimension of $\psi(X)$ by only keeping n entries of i-th order, for each i varying from 1 to k (these entries are selected at random, with the results insensitive to this selection). This yields a reduced $\psi(X)$ dimension of nk, simplifying computations. Therefore the dimensionality of $\psi(X)$ grows only linearly with n and k. While simplifying computations, and reducing the dimension of A that must be learned, we still map X to a higher-dimensional space via $\psi(X)$, as the first terms in $\psi(X)$ are the components of X iteself.

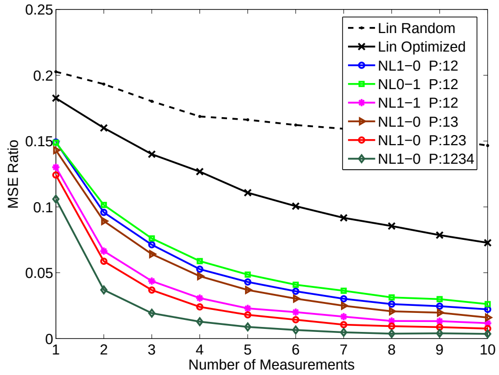

Figure 1. MSE ratio for the nonlinear compressive sensing of GMM data. NLα-β denotes the existence of pure terms X_i^j for $\alpha = 1$ and cross terms $X_i^k \dots X_j^l$ for $\beta = 1$ in the polynomial expansion $\psi(X)$. $P = [p_1 p_2, \dots, p_k]$ represents the contributing powers.

5.1. Signal Recovery

Using simulated GMM data, in Figure 1 the MSE performance of the proposed method, for various polynomial expansions of $\psi(X)$, is compared to linear CS. We consider random and optimized measurement matrices, at the same number of measurements and noise levels. The data is sampled from a GMM distribution $\sum_{t=1}^{T} \pi_t \mathcal{N}(\mu_t, \Sigma_t)$, where the number of GMM components is arbitrarily set to $T = 4$ for demonstration, and the dimension of the input data is $n = 10$. The mean μ_t is a $n \times 1$ vector whose elements are drawn from a uniform distribution on interval $[-1, 1]$. The variance matrix Σ_t is a matrix drawn from a Wishart distribution formed as $\Sigma_t = G_t G_t^T$, where the entities of G are i.i.d. unit-variance zero-mean Gaussian. Measurement noise with variance matrix $\sigma^2 I_{m \times m}$ has been added where $I_{m \times m}$ is $m \times m$ identity matrix, and the SNR defined as $\frac{\mathbb{E}[\Phi(X)^T \Phi(X)]}{m\sigma^2}$ is set to 0 db. We here consider $N_i = 100$ mixture components for each class to estimate $P_{\psi(X)}(x)$. The recovered signal is manifested via the previously defined MMSE estimator.

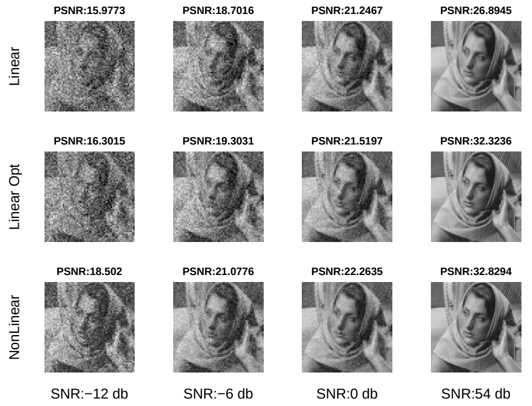

Figure 2. Performance of Bayesian estimation using nonlinear compressed measurements of Image data. The nonlinear transformed data is modeled as GMM with 100 components. A second order polynomial expansion $\psi(X) = [X_1, X_2, \dots X_1^2, \dots X_i X_j, \dots X_n^2]^T$ is used. The rows are corresponding to the i) linear projection with random measurement matrix, ii) optimized linear projection using MIM-GD method (Chen et al., 2012) and iii) the proposed nonlinear optimized measurement, respectively.

In this figure, the notation “Lin Random”, “Lin Optimized”, and “NL” correspond to, respectively, i) linear CS with random measurement matrix with i.i.d Gaussian elements, ii) linear CS with optimized measurement matrix (Carson et al., 2012), and iii) nonlinear CS with optimized (proposed method) measurement matrix. The parameter $P = [p_1 p_2, \dots, p_k]$ in Figure 1 represents the existence of various degrees in the polynomial expansion $\psi(X)$. The terms α and β in (NL-α-β) refer to the existence of pure (X_i^j) and cross terms $X_i^k \dots X_j^l$, respectively. The projection matrix A for the nonlinear CS is optimized based on Theorem 1, using gradient descent to maximize the mutual information between the input and observation vectors. The gradient descent step size and the number of iterations are set to 0.01 and 2000, respectively. The same energy constraint ($E = 1$, arbitrarily) has been applied to all methods considered here.

It is evident from Figure 1 that by adding an intermediate nonlinear mapping stage, the estimation accuracy is significantly improved. The MSE ratio $\frac{\mathbb{E}[\|\hat{X} - X\|^2]}{\mathbb{E}[\|X\|^2]}$ is reduced by a factor of 5 to 10. Higher-order nonlinear mapping provides a better separability of the GMM components, even through its compressed measurements that leads to a better point estimation result. It is noticed that the more higher order nonlinear terms are used, the higher estimation accuracy is obtained. However, this effect saturates and the second or third order polynomials are appropriate choices to avoid unnecessary computational costs, especially when the dimensionality of the input data is fairly large. The results suggest that using only second order expansion with double dimensionality ($r = 2n$) provides a significant signal recovery gain.

Similar result are obtained for the image data recovery as depicted in Figure 2. In this case, the image data is modeled with learned GMM distribution. We use a training set consisting of 500 *jpeg* images from the Berkeley Segmentation Datasets (Carson et al., 2012); this training data is completely distinct from the test data. Each image is split into 4×4 patches. Then, we randomly choose 200 patches from each image and vectorize them to yield $100,000$ vectors of dimension 16. A GMM model with 20 and 100 components are trained for X and $\psi(X)$, respectively. It has been observed that introducing more GMM components for X does not provide a better signal recovery in linear CS case. The results in Figure 2 demonstrate that using nonlinear CS in real applications with high compression requirements provides a promising performance improvement, particularly at low SNR (many similar results were obtained, omitted for brevity).

5.2. Classification

In analogy to the estimation case considered above, classification accuracy can also be improved by incorporating nonlinearity to the input vector based on the same reasoning that a higher dimensional nonlinear space provides more separability, especially when the data classes are not linearly separable or severely corrupted by noise. In other words, the noisy version of $\psi(X)$ with additional nonlinear terms represents richer information about the latent class variable C than the noisy X does. The measurement design employs Theorem 2, maximizing the mutual information between the class labels and the measurements. The aforementioned Bayesian classifier is utilized on the optimized nonlinear compressed version of data, and is compared to the Bayesian classifier applied to the linearly compressed data (the same type classifier is used in all experiments). In addition to the random design of linear measurements, we compare our method to various linear projection design methods, such as the Fisher's Linear discriminant analysis (LDA), where the ratio of the inter-class scattering to the within class scattering of the compressed data is maximized (Fisher, 1936). We also compare the results to linear CS with projections optimized using state-of-the-art information theoretic methods, including information discriminant analysis (IDA) (Nenadic, 2007), Quadratic Renyi (Hild et al., 2006), and the mutual information maximization using gradient descent method (MIM-GD)(Chen et al., 2012). These are comparisons to the very best designed linear measurements in the literature.

A GMM representation with N_i components is trained for each data class $i \in \{1, 2, \ldots, T\}$, resulting a mixture of GMM for $\psi(X)$ (the value of N_i is data- and class-dependent, and N_i is learned, as discussed above). All the competing methods we considered here are derived for the same mixture models, yielding a direct and meaningful comparison.

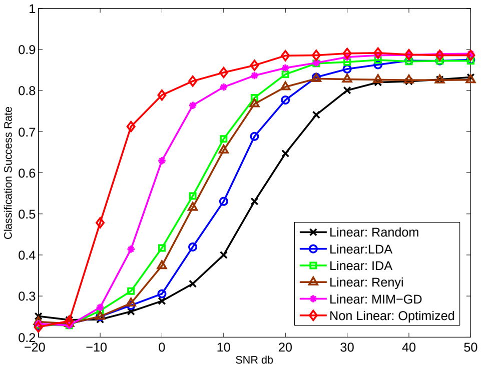

Figure 3. Bayes classification accuracy of GMM data: Nonlinear CS is compared to the linear CS with random and optimized measurement matrices using different methods. Dataset is 6-class Satellite data with training size 4435, testing size 2000 and parameters ($n = 36, N_i = 10, m = 4$).

We consider three datasets including Satellite data, Letter and UPSP digit data, following (Chen et al., 2012), wherein an explanation of the datasets is provided. The results for the classification success rate, averaged over the whole testing set, are presented in Table 1, and Figures 3 and 4.

The results for the linear CS in Table 1 are consistent with the well-know presumption that the information theoretic optimization methods including IDA (Nenadic, 2007), Renyi (Carson et al., 2012) and MIM-GD (Chen et al., 2012) outperforms the correlation-based LDA method, since LDA assumes that the first and second order moments capture the whole data dependence. The results also demonstrate a considerable improvement for the proposed optimized nonlinear CS, with respect to the linear measurements. The gain is markedly superior to the best reported information-theoretic linear projection design (Chen et al., 2012). This, as motivated throughout this paper, is based on the philosophy of providing higher separability by mapping to a higher order nonlinear subspace, as is indirectly used in SVM kernel methods (Bishop & Nasrabadi, 2006). Another intuitive justification is that, by increasing the input vector dimensionality, the measurement noise is corresponding to more input terms, and hence the aggregated impact of the noise on each measurement sample Y_i is reduced due to the structured nature of the signal and the random nature of the noise. The closest result to the proposed nonlinear case is MIM-GD, where a similar mutual information optimization is applied to linear CS. Improved MIM-GD results are manifested because the signal is not assumed to be Gaussian (as in IDA (Nenadic, 2007)) and via a more-accurate representation of the Shannon entropy than the empirically obtained approximation of Renyi entropy (Hild et al., 2006). Nevertheless, including the proposed nonlinearity within the measurement yields further performance improvements. It is notable that the projection design for LDA is performed in one itera-

Table 1. Bayes classification accuracy of GMM data: Nonlinear CS is compared to the linear CS with random and optimized measurement matrices using different methods. Dataset is 26-class Letter data with training size 16000, test size 4000 and parameters $(p = 16, N_i = 10, m = 1, 2, \ldots, 8)$. The noise power is 0 db per sample.

	Bayesian Classification						SVM Kernel Method		
m	Random	LDA	IDA	Renyi	MIM-GD	Nonlinear	Linear	Poly	RBF
1	0.0995	0.1293	0.1435	0.1370	0.1792	**0.2003**	0.1353	0.1275	0.1313
2	0.1383	0.2752	0.2868	0.2107	0.3043	**0.4150**	0.2602	0.2652	0.2592
3	0.2198	0.3523	0.3675	0.3033	0.5052	**0.6205**	0.3200	0.3322	0.3140
4	0.3073	0.4477	0.4577	0.3342	0.6567	**0.7405**	0.3957	0.3812	0.3990
5	0.3655	0.4743	0.5360	0.3585	0.7110	**0.8135**	0.4235	0.4040	0.4200
6	0.4193	0.5323	0.5972	0.4407	0.7710	**0.8425**	0.4412	0.4340	0.4680
7	0.4655	0.5623	0.6485	0.4815	0.7965	**0.8708**	0.4482	0.4540	0.5058
8	0.5022	0.5962	0.6805	0.5410	0.8173	**0.8945**	0.4653	0.4798	0.5350

tion, hence much faster than the iterative methods including IDA, Renyi, MIM-GD and the proposed nonlinear CS. Each iteration in the iterative methods involves matrix inversion, therefore at most proportional to the cube of the dimensionality of input data, n^3. This means an increase by a factor of 8 for our choice of second order polynomial expansion, $(k = 2,\ r = kn)$ with respect to the linear case. Note that the measurement design process (linear or nonlinear) is offline and performed once. Signal recovery or classification with the nonlinear measurement is fast, only very slightly more expensive than the fast inversion of the linear measurement.

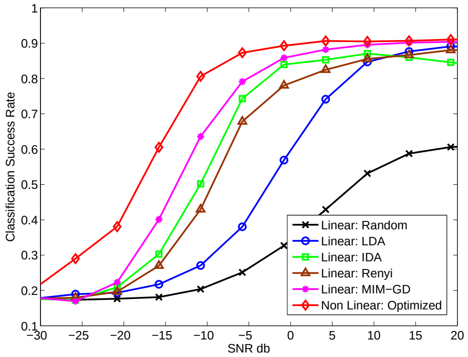

Figure 4. Bayes classification accuracy of GMM data: Nonlinear CS is compared to the linear CS with random and optimized measurement matrices using different methods. Dataset is 10-class Digit data with training size 7291, test size 2007 and parameters $(n = 256, N_i = 1, m = 8)$.

For completeness of comparisons, we also performed kernel SVM classification on the compressed data, similar to (Calderbank & Jafarpour, 2012). We first project the data to the lower dimensional space using random or PCA-based linear projections. We then train an SVM model over the compressed training data and performed classification over the noisy compressed, constituting a fair comparison to our other experiments (equal number of measurements and equal noise level per measurement). We considered an SVM with linear, polynomial of order 3 and RBF kernels, with optimized parameters (Shawe-Taylor & Cristianini, 2004). The results indicate a significant gain for the proposed method over the kernel SVM. Although both kernel SVM and the proposed methods benefit from a similar philosophy of yielding higher separability with nonlinear mapping, the proposed direct mapping of the input to the nonlinear feature space facilitates the optimal measurement design by directly maximizing the mutual information between the class labels and the compressed measurements.

Similar results are obtained for the Satellite and Digit datasets, as shown in Figures 3 and 4. The results in Figure 4 emphasizes the importance of projection design over random projection, when $m \ll n$. These results confirm that the information-theoretic optimized nonlinear sensing outperforms the best reported linear projection design, with a significant margin at low SNR. At high SNR, however, the results are essentially equivalent to the MIM-GD, which is the linear counterpart of the proposed method.

6. Conclusions

We have developed an information-theory-based framework for optimizing nonlinear compressive measurements. Classification and signal-recovery tasks have been addressed, based on new theory for the gradient of mutual information. Specifically, we have derived closed-form gradient-of-mutual-information results, and the optimization has leveraged gradient descent based on this theory. The method is applicable to general source signals, and in the experiments the GMM has been used, with closed-form estimations. Encouraging results for the nonlinear measurement model have been demonstrated on real datasets. It has been demonstrated that the proposed method achieves generally better performance than the best linear measurements and than the nonlinear SVM. The biggest gains over linear methods occur at low SNR, and at high SNR the optimized linear and nonlinear results are similar.

Acknowledgment

The research reported here was funded in part by ARO, DARPA, DOE, NGA and ONR.

References

Aizerman, A, Braverman, E., and Rozoner, L. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 1964.

Bishop, C. and Nasrabadi, N. *Pattern Recognition and Machine Learning*, volume 1. Springer New York, 2006.

Blumensath, T. Compressed sensing with nonlinear observations and related nonlinear optimization problems. *IEEE Transactions on Information Theory*, 2013.

Calderbank, R. and Jafarpour, S. Finding needles in compressed haystacks. In *ICASSP*. IEEE, 2012.

Candès, E.J. and Wakin, M.B. An introduction to compressive sampling. *IEEE Signal Processing Mag.*, 2008.

Candès, E.J., Romberg, J., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure App. Math.*, 59, 2006.

Carson, W.R., Chen, M., Rodrigues, M.R.D., Calderbank, R., and Carin, L. Communications-inspired projection design with application to compressive sensing. *SIAM J. Imaging Sciences*, 5(4), 2012.

Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D., and Carin, L. Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Trans. SP*, 2010.

Chen, M., Carson, W., Rodrigues, M., Calderbank, R., and Carin, L. Communications inspired linear discriminant analysis. In *ICML*, 2012.

Cover, T.M. and Thomas, J.A. *Elements of Information Theory*. Wiley, New York, 2006.

Fisher, R. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

Folland, G.B. *Real Analysis: Modern Techniques and Their Applications*. Wiley New York, 1999.

Guo, D., Shamai, S., and Verdú, S. Additive non-Gaussian noise channels: Mutual information and conditional mean estimation. In *ISIT*, 2005a.

Guo, D., Shamai, S., and Verdú, S. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Trans. IT*, 2005b.

Guo, D., Shamai, S., and Verdú, S. Mutual information and conditional mean estimation in Poisson channels. *IEEE Trans. IT*, 54(5), 2008.

Hellman, M. and Raviv, J. Probability of error, equivocation, and the Chernoff bound. *IEEE Trans. IT*, 1970.

Hild, K.E., Erdogmus, D., Torkkola, K., and Principe, J.C. Feature extraction using information-theoretic learning. *IEEE Trans. PAMI*, 28(9), 2006.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. What is the best multi-stage architecture for object recognition? In *ICCV*, 2009.

Ji, S., Xue, Y., and Carin, L. Bayesian compressive sensing. *IEEE Trans. Signal Processing*, 2008.

Karklin, Y. and Simoncelli, E. Efficient coding of natural images with a population of noisy linear-nonlinear neurons. In *NIPS*, 2011.

Kaski, S. and Peltonen, J. Informative discriminant analysis. In *ICML*, 2003.

Nenadic, Z. Information discriminant analysis: Feature extraction with an information-theoretic objective. *IEEE Trans. PAMI*, 2007.

Palomar, D.P. and Verdú, S. Representation of mutual information via input estimates. *IEEE Trans. IT*, 2007.

Prasad, S. Certain relations between mutual information and fidelity of statistical estimation. *http://arxiv.org/pdf/1010.1508v1.pdf*, 2012.

Schölkopf, B., Smola, A., and Müller, K. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 1998.

Seeger, M.W. and Nickisch, H. Compressed sensing and Bayesian experimental design. In *ICML*, 2008.

Shawe-Taylor, J. and Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge university press, 2004.

Song, L., Smola, A., Borgwardt, K., and Gretton, A. Colored maximum variance unfolding. In *NIPS*, 2008.

Tenenbaum, J., De Silva, V., and Langford, J. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000.

Wang, L., Carlson, D., Rodrigues, M., Wilcox, D., Calderbank, R., and Carin, L. Designed measurements for vector count data. In *NIPS*, 2013.

Wang, L., Carlson, D., Rodrigues, M., Calderbank, R., and Carin, L. A Bregman matrix and the gradient of mutual information for vector Poisson and Gaussian channels. *IEEE Trans. IT, to appear*, 2014.

Xu, W., Wang, M., Cai, J.-F., and Tang, A. Sparse error correction from nonlinear measurements with applications in bad data detection for power networks. *IEEE Trans. Signal Processing*, 2013.