Non-linear Information-theoretic Compressive Projection Design

Abstract

We investigate the projection design problem for non-linear measurement model via the information-theoretic criterion, of interest for non-linear feature design and compressive measurements. Instead of utilizing the popular kernel trick, a direct parametric model on the non-linear map is considered, via which numerous previous kernel methods can be readily manifested. Various gradient of mutual information results for arbitrary input statistics are derived and a gradient descent method is applied. The proposed method of direct non-linear mapping to feature space facilitates the optimal compressive sampling from the feature space and presents a better performance than its linear counterpart. We demonstrate its superior performance on real datasets, for both signal recovery and classification problems.

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 popularities due to its relatively simple formulation and theoretical analysis (Candès et al., 2006), and it is achieved by multiplying a signal of interest with a measurement matrix, and the number of rows of this matrix is small relative to the dimension of the original data vector, where a Gaussian additive measurement noise is assumed (Carson et al., 2012; Ji et al., 2008). However, it has been observed that various sensing systems are intrinsically non-linear, hence the measurements obtained from those systems can only be accurately characterized via non-linear measurement functions (Jarrett et al., 2009; Karklin & Simoncelli, 2011; Xu et al., 2013). Furthermore, compared to its linear counterpart, the non-linear dimensionality reduction techniques have exhibited better performances and flexibilities (Schölkopf et al., 1998; Tenenbaum et al., 2000; Song et al., 2008).

Numerous existing non-linear 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. The original data is believed to be not linearly separable, i.e., one may not find a hyperplane to separate individual clusters. Hence, a non-linear function which maps the data to a new feature space of a much higher dimension than the original one is applied, such that the transformed data becomes linearly separable in that space, in which methods such as Principal Component Analysis (PCA) or Support Vector Machine (SVM) can be readily employed (Shawe-Taylor & Cristianini, 2004). By avoiding a direct computation of inner product in the new space where the knowledge of the specific mapping function is required, the kernel trick suggests that one may calculate the inner product in the new space directly from the original space via a Mercer's kernel (Aizerman et al., 1964).

The kernel trick significantly simplifies the computations involving the non-linear 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 guarantee the performance. In case that a special non-linear space structure is presented (Karklin & Simoncelli, 2011), the kernel trick cannot be easily applied to guarantee such a structural constraint. It is also noted that one may need to store numerous inner product pairs, which may result in a scalability issue for a large number of data points (Shawe-Taylor & Cristianini, 2004) and difficulties for on-line version of the algorithms. Moreover, the kernel methods heavily rely on the fact that inner product is the only information required for the implicit feature space, which prevents further manipulations other than doing classification and analyses of the non-linear map.

Another approach for non-linear dimensionality reduction is to directly model the non-linear compressive function, and design it in the way that the compressed data maximally conveys the original features of interests under some suitable criterion. Among various criteria on measuring the information loss brought by the compressive measurement, mutual information is widely utilized (Hild et al., 2006;

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Kaski & Peltonen, 2003), due to its ubiquitous presence as an information loss measure, as well as its close relationship to Bayesian classification error (Nenadic, 2007). The information-theoretic linear projection design for the Gaussian 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 strategy for the Poisson model has been leveraged in (Wang et al., 2013), where the designed linear projection has been demonstrated to yield improved performance relative to the random choice.

In this paper, we investigate the non-linear measurement design based on the information-theoretic criterion, where the non-linear measurement function is directly modeled after an expansion under certain basis (Taylor's expansion is such an example) and the associated weight parameters are optimized via maximizing the mutual information between the input data and the compressed measurement. Rather than utilizing the kernel trick, one directly optimizes the non-linear function and an explicit expression of the non-linear function is obtained, which may facilitate further analysis and computation. Tasks other than classification such as signal recovery can be readily performed as well under the proposed framework. Undermined by the inability to express the mutual information analytically, gradient-descent methods may be used to optimize the non-linear projection function, where the gradient for non-linear measurement model not only possesses a relatively simple form, but also has connections to the Minimum Mean Square Error (MMSE) estimator.

We present the theoretical results on gradients of mutual information under the non-linear measurement model for both the signal recovery and classification problem, extending previous results under linear measurement model (Guo et al., 2008; Carson et al., 2012; Chen et al., 2012). The theoretical result is for input data with *arbitrary* mixture distribution 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 various real datasets that designed non-linear projection function 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) domain.

2. Non-linear 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, $P_X = \sum_{i=1}^{T} P_C(C=i)P_{X|C}(X|C=i) = \sum_{C=1}^{T} \pi_C P_{X|C}(X|C=i)$, 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 here. We do assume that P_C and $P_{X|C}$ are known or can be estimated from the training datasets.

The non-linear measurement Y is modeled as

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

where $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ is a non-linear measurement function with $m \ll n$ and $W \sim \mathcal{N}(\mathbf{0}, \Sigma)$ is a Gaussian noise with zero mean and variance 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's expansion of Φ. However, it is noted that our theoretical result is valid for arbitrary basis.

Let us consider a polynomial expansion up to k-th 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 $|\alpha| := \sum_i \alpha_i$. We note that if we set $k = 1$, it reduces to a linear measurement model that

$$\begin{aligned} \Phi(X) &= \begin{bmatrix} \Phi_1(X) \\ \vdots \\ \Phi_m(X) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n a_i^{(1)} X_i \\ \vdots \\ \sum_{i=1}^n a_i^{(m)} X_i \end{bmatrix} \\ &= \begin{bmatrix} a_1^{(1)} & \cdots & a_n^{(1)} \\ \vdots & \ddots & \vdots \\ a_1^{(m)} & \cdots & a_n^{(m)} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}. \end{aligned} \tag{3}$$

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., $A_i = [a_{[1,0,\ldots,0]}^{(i)}, a_{[0,1,\ldots,0]}^{(i)}, \ldots, a_{[0,0,\ldots,k]}^{(i)}]$. 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., $\psi(X) = [X_1, X_2, \ldots, X_n, X_1^2, X_1X_2, \ldots, X_n^k]^T$. In this manner, we can rewrite the non-linear measurement function Φ as

$$\Phi(X) = A\psi(X). \tag{4}$$

Note that the basis vector $\psi(X)$ is fixed once we fix the basis and the expansion order.

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 object is to design Φ to maximize the information content in Y about the class label C and we term this problem as the "classification" problem.

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

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

$$= \arg\max_{A} I(X;Y). \tag{6}$$

In the case for classification problem, we adopt the criterion

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

$$= \arg\max_{A} I(C;Y). \tag{8}$$

We consider all above optimization problems together with the energy constraint

$$\text{tr}[\Phi\Phi^T] \leq E. \tag{9}$$

with E being a constant. The Shannon entropy with natural logarithm is considered throughout.

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

$$\text{MMSE} \geq \frac{1}{2\pi e} \exp\{2[h(Y) - I(X;Y)]\} \tag{10}$$

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 $P_e = \int P_Y(y)[1 - \max_X P_{X|Y}(x|y)]dy$, and noting that it has been shown in (Hellman & Raviv, 1970) that

$$P_e \leq \frac{1}{2} H(X|Y) \tag{11}$$

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

2.2. Connection to Kernel Methods

Our approach differs from the kernel methods in an apparent way that, rather than implicitly transforming the data to a high dimensional feature space, it directly compresses the data to a lower dimensional space. However, the proposed method also possesses an interpretation, in which a much higher dimensional space is involved. To be more specific, according to (4), it can be recognized that the data X is first transformed to a much higher dimensional space via the basis vector $\psi(X)$, where a linear projection A is applied afterwards. In other words, another perspective to understand the non-linear projection function is that it is equivalent to a linear projection which maps from a very high dimensional space spanned by the non-linear basis ψ to a lower dimensional space.

As a matter of fact, 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 $\phi(x) : \mathbb{R}^n \to \mathbb{R}^l$ with $l \gg n$ be the implicit feature mapping used in the kernel methods. By 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 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 the projection matrix P is constituted indirectly via the kernel trick, by relaxing PA_ϕ to one matrix A, it only requires 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 (order of the Taylor's expansion in this case).

2.3. Connection to Linear Compressive Sensing

A vast amount of efforts has been focused on investigation of the recoverability for the linear compressive measurement models (Candès & Wakin, 2008). It is straightforward to pose a similar question that whether the pure non-linear map (excluding linear map) is able to achieve a similar or even better performance.

As discussed in previous section, it is possible to view the non-linear projection $\Phi(x)$ as a linear projection which maps from a space spanned by the non-linear basis $\psi(X)$, i.e., $\Phi(x) = A\psi(X)$. Recall that the Johnson-Lindenstrauss (JL) Lemma (Johnson & Lindenstrauss, 1984) essentially claims the existence of a linear map from which the linearly compressed measurements preserve almost all clustering information, provided that some technical conditions are satisfied. It is easy to see that for any two points $x, y \in \mathbb{R}^n$, $\|\psi(X) - \psi(y)\|_2 \geq \|x - y\|_2$ as ψ being

the polynomial basis, which implies that the inter-cluster distance will not be decreased when the data is transformed via the basis vector ψ. By constituting A via a JL matrix, the non-linear measurement is able to preserve all the original clustering information contained in the data as well, Therefore a non-linear map is guaranteed to at least match the performance under a linear map.

With a performance guarantee for the non-linear projection via the JL Lemma, we now argue that the non-linear projection can indeed provide a better performance. In (Calderbank & Jafarpour, 2012), a random linear projection has been constituted in the way that, with high probability, the performance of the linear SVM in the linearly compressed domain is closed to performing that in the original data domain. When the data X is non-linearly transformed to a high dimension space spanned by $\psi(X)$ where a linear separation is possible, it has been widely recognized that methods such as linear SVM applied in the transformed domain will consistently achieve a much better performance than directly carried out in the original data domain, and such a performance gain is inherited by constituting A as the random matrix specified in (Calderbank & Jafarpour, 2012).

Based on previous discussions, it strongly suggests that the non-linear compressive sensing can potentially attain a better classification performance than its linear counterpart. While for signal recovery problem, a better performance under the non-linear projection is guaranteed provided that the Taylor's expansion of the non-linear projection satisfies properties related to the Restricted Isometry Property (RIP) (Blumensath, 2013)

Most of previously mentioned performance guarantees are built on randomized choices of the non-linear projection. Alternatively, one may design the non-linear map with the goal of maximizing mutual information as specified in (6) and (8) via a gradient descent method described as follow.

3. Gradients of Mutual Information for Non-linear Measurement Model

3.1. Explicit Gradient Formulas

The mutual information terms in (6) and (8) generally do not possess any known analytical form. Hence, the gradient of mutual information with respect to A may shed light on solving those optimization problems, and it may be used in numerical experiments. It is therefore 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 non-linear 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 $I(X;Y)$ for the non-linear measurement model in (1) can be expressed as*

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

More generally, if $\Phi(X) = A\mathcal{K}(X)$ where $\mathcal{K}(X)$ is an arbitrary functional basis vector (e.g. Chebyshev or Hermite polynomials) and A is the corresponding coefficient matrix associated with that basis vector $\mathcal{K}(X)$, we have

$$\nabla_A I(X;Y) = \Sigma^{-1} A \mathbb{E}[[\mathcal{K}(X) - \mathbb{E}[\mathcal{K}(X)|Y]][\mathcal{K}(X) - \mathbb{E}[\mathcal{K}(X)|Y]]^T]. \tag{13}$$

The above theorem generalizes the scalar result for non-linear 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*

$$\nabla_A I(X;Y) = \Sigma^{-1} A \mathbb{E}[[X - \mathbb{E}[X|Y]][X - \mathbb{E}[X|Y]]^T]. \tag{14}$$

The gradient of mutual information between 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 non-linear measurement model in (1) can be expressed as*

$$\nabla_A I(C;Y) = \Sigma^{-1} A \times \mathbb{E}[[\mathbb{E}[\psi(X)|Y,C] - \mathbb{E}[\psi(X)|Y]][\mathbb{E}[\psi(X)|Y,C] - \mathbb{E}[\psi(X)|Y]]^T]. \tag{15}$$

More generally, if $\Phi(X) = A\mathcal{K}(X)$ where $\mathcal{K}(X)$ is an arbitrary functional basis vector and A is the corresponding coefficient matrix associated with that basis vector $\mathcal{K}(X)$, we have

$$\nabla_A I(C;Y) = \Sigma^{-1} A \times \mathbb{E}[[\mathbb{E}[\mathcal{K}(X)|Y,C] - \mathbb{E}[\mathcal{K}(X)|Y]][\mathbb{E}[\mathcal{K}(X)|Y,C] - \mathbb{E}[\mathcal{K}(X)|Y]]^T]. \tag{16}$$

The proofs of above theorems are presented in the Supplementary Material. Note that Theorems 1 and 2 are valid for *arbitrary* P_X or $P_{X|C}$ provided that the regularity conditions are satisfied.

3.2. Gradient-based Numerical Design

A numerical solution to the optimizations in (6) and (8) 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; 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 interests and proj$(\cdot)$ projects the matrix to the feasible set defined by the energy constraint in (9), *i.e.*, re-normalize 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. Experiments

In this section, we examine the proposed non-linear Compressive Sensing (CS) with optimized measurement design for both signal recovery and classification problems. The gradient results in Theorem 1 and 2 are valid for arbitrary mixture distribution of X and can be readily performed via the Monte Carlo integration, provided the posterior $P_{\psi(X)|Y}$ is easily sampled. In order to estimate the distribution of $\psi(X)$, we employ the kernel density estimation (KDE) (Silverman, 1986) on N transformed data samples $\{\psi(X_i)\}_{i=1}^N$ which are directly obtained by applying the map $\psi(X)$ on the training data samples $\{X_i\}_{i=1}^N$, and the estimated density can be expressed as

$$\hat{P}_{\psi(X)}(x) = \frac{1}{N}\sum_{i=1}^{N} K(\frac{x - \psi(X_i)}{h}), \qquad (17)$$

where $K(\cdot)$ is the kernel function and $h > 0$ is the bandwidth parameter which can be set empirically via Silverman's criterion (Silverman, 1986). In the paper, the kernel function $K(\cdot)$ is selected as the Gaussian kernel (also known as the RBF kernel). Similar estimator can be applied to estimate $P_{X|C}$ as well. Via the RBF kernel, it can be readily recognized that the estimated density functions $P_\psi(X)$ and P_X are manifested as Gaussian Mixture Models (GMM), from which samples can be drawn easily.

In addition to being easily sampled, the GMM type estimate has the advantage of an analytic posterior $P_{X|Y}$, which is also a GMM (Chen et al., 2010), specifically, we have that $P_{X|Y} = \sum_{i=1}^T \tilde{\pi}_i P_{X|Y,C=i}$, with analytic expressions for $\{\tilde{\pi}_i\}$ and $P_{X|Y,X=i}$ derived in (Chen et al., 2010). This posterior naturally induces a Bayesian classifier $\max_i P_{C=i|Y}$, where $P_{X=i|\tilde{Y}} = \tilde{\pi}_i$, and a MMSE estimator $\tilde{X}$ for the input data X, $\tilde{X} := \mathbb{E}[X|Y]$.

In order to obtain better results and faster convergence, we take the following considerations in forming the basis vector $\psi(X)$. When mapping the input vector X to the nonlinear feature space with dimensionality r using polynomial expansion of order k ($X \to \psi'(X)$), we use $\psi'(X) = [\gamma_1 X_1, \gamma_2 X_2, \ldots, \gamma_r X_n^k]^T$. The coefficients $\gamma_i = n\mathbb{E}[[\psi(X)]_j^2]/\mathbb{E}[X^T X]$ are used to normalize the basis vector through balancing the contributing nonlinear terms of different degrees and avoid dominating various terms one another. This coefficients can be found by calculating the moments of polynomial expansion of Gaussian terms using *generalized Hermite polynomials* or empirically obtained by sampling the nonlinear expansion of learned GMM model. The energy constraint in (9) with this normalization reduces to $\text{tr}[AA^T] = 1$, where $E = x^T x$.

To avoid extra computational cost due to the extremely large dimensional $\psi'(X)$, we randomly choose r rows by $\psi(X) = \Pi\psi'(X)$, where Π is an identity matrix with some of the rows deleted in random. In all simulations, we choose to keep n terms of each power, hence the dimensionality of $\psi(X)$ grows only linearly with n and k. In sequel, we present some results obtained for both Bayesian signal recovery and classification problems using the optimized nonlinear compressed sensing and GMM data.

4.1. Signal Recovery

In Figure 1, the performance of the proposed method for various polynomial expansions using simulated GMM data is compared to the linear CS with 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 $T = 4$ and the dimensionality of the input data is $n = 10$. The measurement noise is set to 0 db. We use KDE with RBF kernel and 100 training points.

In this figure, the notation "Lin Random", "Lin Optimized", and "NL" correspond to i) linear CS with random measurement matrix with i.i.d Gaussian terms, ii) linear CS with optimized measurement matrix and iii) Non linear CS with optimized measurement matrix, respectively. The optimization method for linear CS is based on the mutual information maximization proposed in (Carson et al., 2012). The parameter $P = [P_1 P_2, \ldots, P_k]$ in Figure 1 represents the existence of various degrees in the polynomial expan-

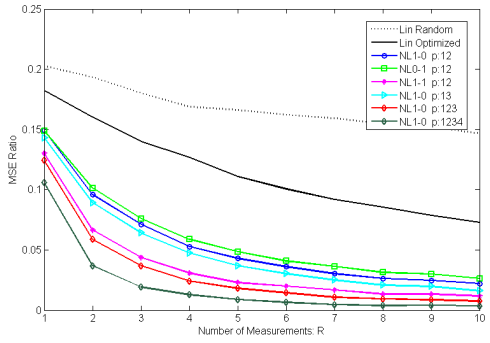

Figure 1. MSE ratio for nonlinear compressive sensing of GMM data. NL$\alpha-\beta$ denotes the existence of pure terms x_i^j for $\alpha=1$ and cross terms $x_i^k...x_j^l$ for $\beta=1$ in the polynomial expansion. $P=[i,j,..k]$ represents the contributing powers.

Figure 2. Performance of Bayesian estimation using nonlinear compressive sensing of Image data. A second order polynomial expansion $\psi(X)=[X_1,X_2,\dots X_1^2,...X_iX_j,\dots X_n^2]^T$ is used.

sion $\psi(X)$. The terms α and β in $(NL-\alpha-\beta)$ refer to the existence of pure (X_i^j) and cross terms, respectively. The projection matrix A for nonlinear CS is optimized based on Theorem 1 using gradient descent method to maximize the mutual information between the input and observation vectors. Gradient descent step and the number of iterations are set to 0.01 and 2000, respectively.

It is evident from Figure 1 that by adding an intermediate nonlinear mapping, the estimation accuracy is significantly improved. The MSE ratio $\frac{\mathbb{E}[||\hat{X}-X||]^2}{\mathbb{E}[||X||]^2}$ is reduced with 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 results. It is noticed that the more of 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. It is also observed that for the second order polynomial expansion, using mixture of both pure and cross terms are advantageous, while at higher order polynomials, pure terms present a better performance. This flexibility of choosing various non-linear terms is not easy to obtain in SVM kernel trick. 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 GMM distribution. We use a training set consist of 500 *jpeg* images from the Berkeley Segmentation Datasets (Ber). 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 $T=20$ components is trained using EM method. Hence, a GMM prior is considered for Bayesian estimation of the test data. The results in this figure demonstrate that using nonlinear CS in real applications with high compression requirements provides a promising performance improvement, specially at low SNR regime. The ultimate number and energy of measurements for both linear and non-linear cases are equal and the only cost is a higher complexity, which is affordable when using only second order terms.

4.2. Classification

In analogy to the estimation scenario, the classification accuracy can be improved by incorporating nonlinearity to the input vector based on the same reasoning that a higher dimensional non-linear space provides more separability, especially when the data classes are not linearly separable or severely corrupted by noise. The projection matrix design lays on Theorem 2 that maximizes the mutual information between the class labels and the measurements. A Bayesian classification is performed on the optimized nonlinear compressed version of data and is compared to the classification of linearly compressed data using the state of the art projection designs, including LDA, IDA, Quadratic Renyi, and Chen's scheme, with details presented in (Chen et al., 2012) and references therein.

A GMM model with T components is trained for each data class c, resulting a mixture of GMM for the input data. Then, Bayesian classification is applied to the nonlinear compressed measurements. We use three datasets including Satellite data, Letter and UPSP digit data following (Chen et al., 2012), wherein the explanation of datasets is

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 Letter data with training data size 16000 and parameters ($p = 16, c = 26, T = 10, m = 8$).

CLASSIFICATION:	BAYESIAN COMPRESSIVE SENSING						SVM KERNEL METHOD		
PROJ. DESIGN:	RANDOM	LDA	IDA	RENY	CHEN	NONLINEAR	LINEAR	POLYNOMIAL	RBF
M=1	0.0995	0.1293	0.1435	0.1370	0.1792	**0.2003**	0.1353	0.1275	0.1313
M=2	0.1383	0.2752	0.2868	0.2107	0.3043	**0.4150**	0.2602	0.2652	0.2592
M=3	0.2198	0.3523	0.3675	0.3033	0.5052	**0.6205**	0.3200	0.3322	0.3140
M=4	0.3073	0.4477	0.4577	0.3342	0.6567	**0.7405**	0.3957	0.3812	0.3990
M=5	0.3655	0.4743	0.5360	0.3585	0.7110	**0.8135**	0.4235	0.4040	0.4200
M=6	0.4193	0.5323	0.5972	0.4407	0.7710	**0.8425**	0.4412	0.4340	0.4680
M=7	0.4655	0.5623	0.6485	0.4815	0.7965	**0.8708**	0.4482	0.4540	0.5058
M=8	0.5022	0.5962	0.6805	0.5410	0.8173	**0.8945**	0.4653	0.4798	0.5350

son setup (equal number of measurements and equal noise level). We tried SVM with linear, polynomial and RBF kernels with optimized parameters. The results present a significant gain for the proposed method over the SVM kernel. The obvious reason is that although both SVM kernel and the proposed methods benefit from the similar philosophy of proving separability with nonlinear mapping, the proposed direct mapping of the input to the nonlinear feature space, facilitates the optimal projection design to maximize the mutual information between the class labels and the compressed measurements.

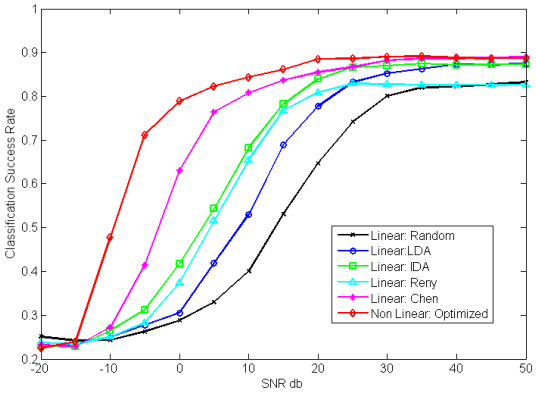

Figure 4. Bayes classification accuracy of GMM data: Nonlinear CS is compared to the linear CS with random and optimized measurement matrices using different methods. Dataset is Satellite data with training data size 4435 and parameters ($p = 36, c = 6, T = 10, m = 4$).

Similar results are obtained for the Satellite and Digit datasets that as shown in Figures 4 and 5. The results in Figure 5 emphasizes on the importance of projection design, when the number of compressive measurements with respect to the data dimensionality is relatively low. 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 regime, however, the results are equivalent to the Chen's information-theoretic optimization method, which is the linear counterpart of the proposed method.

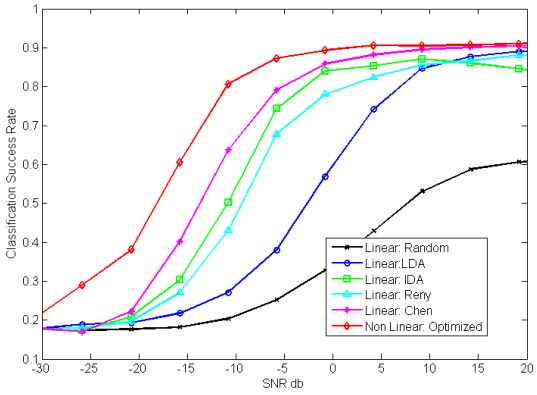

Figure 5. Bayes classification accuracy of GMM data: Nonlinear CS is compared to the linear CS with random and optimized measurement matrices using different methods. Dataset is Digit data with training data size 7291 and parameters ($p = 256, c = 10, T = 1, m = 8$).

5. Conclusions

We have developed a theory for optimization of the nonlinear projection for the non-linear measurement model, based on maximizing the mutual information for signal recovery or classification. Albeit manifested in different fashion, numerous previous kernel methods can be unified

via the proposed model. Tasks including classification and signal recovery have been taken into account under the proposed framework. We have derived various gradient of mutual information results, and the optimization has been leveraged via a gradient descent method, which is valid for arbitrary mixture model of the input distribution. Further, rather than assuming specific input probability model, a kernel density estimation has been employed to capture the true underlying distribution as well as simplifying the involved Monte Carlo integration. Encouraging results for the non-linear measurement model have been demonstrated on real datasets, and it has been shown that the proposed method has achieved a generally better performance, compared to various linear methods and the non-linear SVM.

References

Berkeley segmentation datasets. URL http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/resources.html.

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.*, 25(2), 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 TSP*, 2010.

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

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*. IEEE, 2005a.

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

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

Hellman, M. and Raviv, J. Probability of error, equivocation, and the Chernoff bound. *IEEE Trans. on Info. Theory*, 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.

Johnson, W. and Lindenstrauss, J. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26(189-206):1, 1984.

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 TPAMI*, 2007.

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

Silverman, B. *Density Estimation for Statistics and Data Analysis*, volume 26. CRC press, 1986.

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 Transactions on Information Theory, 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.

Nonlinear Information-theoretic Compressive Projection Design

1 Proof of Theorem 1

Proof of Theorem 1. The gradient of mutual information $\nabla_A I(X;Y)$ can be expressed as

$$\nabla_A I(X;Y) = \nabla_A [h(Y) - h(Y|X)] \tag{1}$$
$$= \nabla_A h(Y) \tag{2}$$

The equality follows from the fact that $h(Y|X) = h(\mathcal{N}(A\psi(X), \Sigma))$, which is a constant with respect to A. Hence, we have

$$\nabla_A I(X;Y) = \nabla_A h(Y) \tag{3}$$
$$= -\nabla_A \int \log P_Y(y) P_Y(y) dy \tag{4}$$
$$= -\int \nabla_A P_Y(y) dy - \int \log P_Y(y) \nabla_A P_Y(y) dy \tag{5}$$
$$= -\nabla_A \int P_Y(y) dy - \int \log P_Y(y) \nabla_A P_Y(y) dy \tag{6}$$
$$= -\int \log P_Y(y) \nabla_A P_Y(y) dy, \tag{7}$$

where the change of operator orders in (5) and (6) follows from the assumed regularity conditions. We now calculate the term $\nabla_A P_Y(y)$

$$\nabla_A P_Y(y) = \nabla_A \int P_{Y|X}(y|x) P_X(x) dx \tag{8}$$
$$= \int P_X(x) \nabla_A P_{Y|X}(y|x) dx \tag{9}$$
$$= \int P_X(x) \nabla_A \mathcal{N}(y; A\psi(x), \Sigma) dx \tag{10}$$
$$= \int P_X(x) \nabla_y \mathcal{N}(y; A\psi(x), \Sigma) \psi^T(x) dx \tag{11}$$
$$= \int P_X(x) \nabla_y P_{Y|X}(y|x) \psi^T(x) dx. \tag{12}$$

Plug in (12) back to (7) and by the Fubini's Theorem [1], we have

$$\nabla_A I(X;Y) = -\int \log P_Y(y) P_X(x) \nabla_y P_{Y|X}(y|x) \psi^T(x) dxdy \tag{13}$$

$$= -\int \nabla_y (\log P_Y(y)) P_X(x) P_{Y|X}(y|x) \psi^T(x) dxdy \tag{14}$$

$$= -\int \frac{P_X(x) P_{Y|X}(y|x)}{P_Y(y)} \nabla_y P_Y(y) \psi^T(x) dxdy \tag{15}$$

$$= -\int \nabla_y P_Y(y) \mathbb{E}[\psi^T(X)|Y=y] dy, \tag{16}$$

where (14) invokes the integration by parts [1]. $\nabla_y P_Y(y)$ can be expressed as

$$\nabla_y P_Y(y) = \int \nabla_y P_{Y|X}(y|x) P_X(x) dx \tag{17}$$

$$= \int \nabla_y \log P_{Y|X}(y|x) P_{Y|X}(y|x) P_X(x) dx \tag{18}$$

$$= \int \nabla_y \log \mathcal{N}(y; A\psi(x), \Sigma) P_{X|Y=y}(x|y) P_Y(y) dx \tag{19}$$

$$= \mathbb{E}[\nabla_Y \log \mathcal{N}(Y; A\psi(X), \Sigma)|Y=y] P_Y(y). \tag{20}$$

It is straightforward to calculate

$$\nabla_y \log \mathcal{N}(y; A\psi(x), \Sigma) = -\Sigma^{-1}(y - A\psi(x)). \tag{21}$$

Hence, we have

$$\nabla_y P_Y(y) = \mathbb{E}[-\Sigma^{-1}(Y - A\psi(X))] P_Y(y) \tag{22}$$

$$= -\Sigma^{-1} \mathbb{E}[W|Y=y] P_Y(y). \tag{23}$$

Combining (23) with (16), we obtain

$$\nabla_A I(X;Y) = \int \Sigma^{-1} \mathbb{E}[W|Y=y] P_Y(y) \mathbb{E}[\psi^T(X)|Y] dy \tag{24}$$

$$= \Sigma^{-1} \mathbb{E}[\mathbb{E}[W|Y] \mathbb{E}[\psi^T(X)|Y]] \tag{25}$$

$$= \Sigma^{-1} \mathbb{E}[\mathbb{E}[Y - A\psi(X)|Y] \mathbb{E}[\psi^T(X)|Y]] \tag{26}$$

$$= \Sigma^{-1} \mathbb{E}[\mathbb{E}[Y\psi^T(X)|Y]] - A\mathbb{E}[\mathbb{E}[\psi(X)|Y] \mathbb{E}[\psi^T(X)|Y]] \tag{27}$$

$$= \Sigma^{-1} \mathbb{E}[\mathbb{E}[(A\psi(X) + W)\psi^T(X)]] - A\mathbb{E}[\mathbb{E}[\psi(X)|Y] \mathbb{E}[\psi^T(X)|Y]] \tag{28}$$

$$= \Sigma^{-1} \mathbb{E}[\mathbb{E}[A\psi(X)\psi^T(X)]] - A\mathbb{E}[\mathbb{E}[\psi(X)|Y] \mathbb{E}[\psi^T(X)|Y]] \tag{29}$$

$$= \Sigma^{-1} A\mathbb{E}[\mathbb{E}[\psi(X)\psi^T(X)]] - \mathbb{E}[\mathbb{E}[\psi(X)|Y] \mathbb{E}[\psi^T(X)|Y]] \tag{30}$$

$$= \Sigma^{-1} A\mathbb{E}[\mathbb{E}[\psi(X)\psi^T(X)]] - \mathbb{E}[\psi(X)\mathbb{E}[\psi^T(X)|Y]$$
$$- \mathbb{E}[\psi(X)|Y]\mathbb{E}[\psi^T(X)]] + \mathbb{E}[\psi(X)|Y]\mathbb{E}[\psi^T(X)|Y] \tag{31}$$

$$= \Sigma^{-1} A\mathbb{E}[[\psi(X) - \mathbb{E}[\psi(X)|Y]][\psi(X) - \mathbb{E}[\psi(X)|Y]]^T]. \tag{32}$$

Note that we do not assume specific form of $\psi(x)$ in previous proof, thus it applies to arbitrary basis vector $\psi(x)$, provided that it satisfies the regularity conditions. □

2 Proof of Theorem 2

Proof of Theorem 2. First we notice that

$$I(C;Y) = h(Y) - h(Y|C) \tag{33}$$
$$= h(Y) - h(Y|X) + h(Y|X,C) - h(Y|C) \tag{34}$$
$$= I(X;Y) - I(X;Y|C), \tag{35}$$

where the second equality is due to the fact that $C \to X \to Y$ forms a Markov chain and $P_{Y|X,C} = P_{Y|X}$. Following by the similar steps in the proof of Theorem 1, we have

$$\nabla_A I(X;Y|C) = \Sigma^{-1} A \mathbb{E}[[\psi(X) - \mathbb{E}[\psi(X)|Y,C]][\psi(X) - \mathbb{E}[\psi(X)|Y,C]]^T]. \tag{36}$$

Let μ denote the counting measurement on $\mathbb{R}$. Using the following facts that

$$\mathbb{E}[\mathbb{E}[\psi(X)|Y,C]\mathbb{E}[\psi(X)^T|Y,C]] = \mathbb{E}[\mathbb{E}[\psi(X)|Y,C]\psi(X)^T] = \mathbb{E}[\psi(X)\mathbb{E}[\psi(X)^T|Y,C]] \tag{37}$$
$$\mathbb{E}[\mathbb{E}[\psi(X)|Y,C]\mathbb{E}[\psi(X)^T|Y]] = \mathbb{E}[\mathbb{E}[\psi(X)|Y]\mathbb{E}[\psi(X)^T|Y,C]] = \mathbb{E}[\mathbb{E}[\psi(X)|Y]\mathbb{E}[\psi(X)^T|Y]] \tag{38}$$
$$\mathbb{E}[\psi(X)\mathbb{E}[\psi(X)^T|Y]] = \mathbb{E}[\mathbb{E}[\psi(X)|Y]\psi(X)^T] = \mathbb{E}[\mathbb{E}[\psi(X)|Y]\mathbb{E}[\psi(X)^T|Y]], \tag{39}$$

we have the following expressions

$$\begin{aligned}
&\mathbb{E}[[\psi(X) - \mathbb{E}[\psi(X)|Y]][\psi(X) - \mathbb{E}[\psi(X)|Y]]^T] \\
&= \mathbb{E}[\psi(X)\psi(X)^T - \psi(X)\mathbb{E}[\psi(X)^T|Y] - \mathbb{E}[\psi(X)|Y]\psi(X) + \mathbb{E}[\psi(X)|Y]\mathbb{E}[\psi(X)^T|Y]] \qquad (40)\\
&= \mathbb{E}[\psi(X)\psi(X)^T - \psi(X)\mathbb{E}[\psi(X)^T|Y,C] - \mathbb{E}[\psi(X)|Y,C]\psi(X) + \mathbb{E}[\psi(X)|Y,C]\mathbb{E}[\psi(X)^T|Y,C] \\
&+ \mathbb{E}[\psi(X)|Y,C]\mathbb{E}[\psi(X)^T|Y,C] - \mathbb{E}[\psi(X)|Y,C]\mathbb{E}[\psi(X)^T|Y] - \mathbb{E}[\psi(X)|Y]\mathbb{E}[\psi(X)^T|Y,C] \\
&+ \mathbb{E}[\psi(X)|Y]\mathbb{E}[\psi(X)^T|Y]] \qquad (41)\\
&= \mathbb{E}[[\psi(X) - \mathbb{E}[\psi(X)|Y,C]][\psi(X) - \mathbb{E}[\psi(X)|Y,C]]^T] \qquad (42)\\
&+ \mathbb{E}[[\mathbb{E}[\psi(X)|Y] - \mathbb{E}[\psi(X)|Y,C]][\mathbb{E}[\psi(X)|Y] - \mathbb{E}[\psi(X)|Y,C]]^T].
\end{aligned}$$

Therefore, we have

$$\nabla_A I(C;Y) = \nabla_A I(X;Y) - \nabla_A I(X;Y|C) \tag{43}$$
$$= \mathbb{E}[[\psi(X) - \mathbb{E}[\psi(X)|Y]][\psi(X) - \mathbb{E}[\psi(X)|Y]]^T]$$
$$- \mathbb{E}[[\psi(X) - \mathbb{E}[\psi(X)|Y,C]][\psi(X) - \mathbb{E}[\psi(X)|Y,C]]^T] \tag{44}$$
$$= \mathbb{E}[[\mathbb{E}[\psi(X)|Y] - \mathbb{E}[\psi(X)|Y,C]][\mathbb{E}[\psi(X)|Y] - \mathbb{E}[\psi(X)|Y,C]]^T]. \tag{45}$$

□

References

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