DIMENSION REDUCTION IN REGRESSION USING GAUSSIAN MIXTURE MODELS

Majid Mirbagheri[1], *Yanbo Xu*[1], *Shihab Shamma*[1,2]

The Institute for Systems Research[1], Department of Electrical and Computer Engineering[2]
University of Maryland College Park
{mbagheri, yanbohsu, sas}@umd.edu

ABSTRACT

Linear–Nonlinear regression models play a fundamental role in characterizing nonlinear systems. In this paper, we propose a method to estimate the linear transform in such models equivalent to a subspace of a small dimension in the input space that is relevant for eliciting response. The novel aspect of this work is the formulation of the mutual information between the transformed inputs and output as a closed-form function of the parameters of their joint density in the form of Gaussian Mixture Models and we subsequently maximize this measure to find relevant dimensions. Instead of a commonly used mutual information measure based on Kullback-Leibler divergence, we use a measure called Quadratic Euclidean Mutual Information. Through experiments on both synthesized data and real MEG recordings, the effectiveness of the proposed method is demonstrated.

Index Terms— dimension reduction, regression, mutual information, gaussian mixture models

1. INTRODUCTION

Nonlinear system identification techniques have long been applied to the study of real systems. Of particular interest are cascade linear–nonlinear (LN) models which have recently been shown to uniformly approximate, to an arbitrary degree of precision, any continuous function [1].

Here, we focus on the estimation of the linear part which usually performs dimension reduction on the inputs of the system, based on the knowledge that the number of dimensions in the input space important for eliciting a response is typically smaller than the size of the input. The transformed input space highlights the "important" feature dimensions, thus simplifying model analysis and design.

The pioneering studies on Sliced Inverse Regression (SIR) [2] and Sliced Average Variance Estimator (SAVE) [3] have drawn attention to the research for dimension reduction in regression which aims at reducing the dimension of a vector-valued predictor $\boldsymbol{X}$ while the regression relation with a real-valued Y is preserved. The proposed method falls into this category when the system output is assumed to be a real random variable. Various methods have been suggested to overcome the well known limitations in SIR and SAVE. To achieve better estimates of $E(\boldsymbol{X}|Y)$ and $E(\boldsymbol{X}\boldsymbol{X}^T|Y)$, recently, a series of methods via maximum likelihood [4] have been suggested, in which the conditional distribution of X given certain Y is assumed to be gaussian. Alternatively, Gaussian Mixture Models (GMM) is adopted to approximate the underlying joint probability distribution of $\boldsymbol{X}$ and Y, avoiding the mismatch between assumptions and real data. Besides, as our primary interest is to analyze cascade linear–nonlinear models, mutual information [5] is brought to measure the relevance of projected $\boldsymbol{X}$ with Y in order to avoid imposing certain functional forms on the nonlinearity of the system. Although the idea of using mutual information to perform dimension reduction is not new, differing from previous methods, we are particularly interested in dimension reduction in nonlinear systems with continuous response. And by combining GMM and Quadratic Mutual Information (explained later), our method provides a closed-form measure of mutual information between the atransformed input and output.

The structure of the paper is as follows. The next section describes the proposed method in details. And experimental results are presented in section 3. And finally we conclude this work with section 4.

2. DIMENSION REDUCTION IN REGRESSION USING GMM

In this section, we describe the proposed method to conduct dimension reduction in regression.

2.1. Quadratic Mutual Information

Based on Shannon's definition, MI between two random variables can be viewed as Kullback-Leibler divergence between their joint density and the product of their marginal densities. We examine now alternative divergence measures for our purpose. In [6], Kapur argued that if the aim is not to calculate an absolute value of the divergence, but rather to find a distribution that minimizes/maximizes the divergence, the axioms used in deriving the measure can be relaxed and yet the result of the optimization is the same distribution. Now if we define

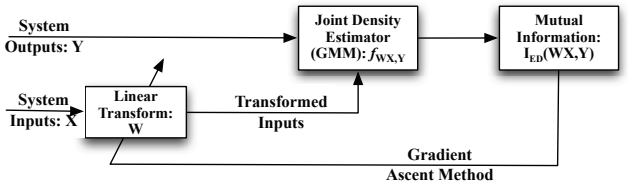

Fig. 1. Learning linear transforms by maximizing the mutual information between system outputs and input projections.

the Euclidean Distances between two distributions f and g as:

$$D_{ED} = \int \big(f(x) - g(x)\big)^2 dx \tag{1}$$

then Quadratic Mutual Information Euclidean Distance (denoted by I_{ED} henceforth) of two random variables X and Y would be defined as the Euclidean distance between their joint pdf and the factorized marginal pdf, and is written as:

$$\begin{aligned} I_{ED}(X,Y) &= \\ &= D_{ED}\big(f_{X,Y}(x,y), f_X(x)f_Y(y)\big) \\ &= \iint \big(f_{X,Y}(x,y) - f_X(x)f_Y(y)\big)^2 dxdy \end{aligned} \tag{2}$$

Now assume that two random vectors X and Y, respectively $n-$ and $m-$dimensional, have a joint density modeled as a Gaussian mixture of the following form:

$$f_Z(\boldsymbol{z}) = \sum_{i=1}^{K} \omega_i G(\boldsymbol{z} - \boldsymbol{\mu}_i, C_i), \tag{3}$$

where Z is the random vector describing their joint behaviors ($Z = \begin{bmatrix} X \\ Y \end{bmatrix}$), ω_i are non-negative weighting coefficients with $\sum_i \omega_i = 1$ and G is the normal distribution with mean $\boldsymbol{\mu}$ and covariance C.

$$G(\boldsymbol{z} - \boldsymbol{\mu}, C) = |(2\pi)^{m+n} C|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T C^{-1}(\boldsymbol{z}-\boldsymbol{\mu})} \tag{4}$$

Notice that X and Y will also have marginal densities of the form:

$$f_X(\boldsymbol{x}) = \sum_{i=1}^{K} \omega_i G(\boldsymbol{x} - \boldsymbol{\mu}_i^X, A_i), \tag{5}$$

$$f_Z(\boldsymbol{y}) = \sum_{i=1}^{K} \omega_i G(\boldsymbol{y} - \boldsymbol{\mu}_i^Y, B_i), \tag{6}$$

with $\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^X \\ \boldsymbol{\mu}_j^Y \end{bmatrix}$ and $C_i = \begin{bmatrix} A_i & D_i \\ D_i^T & B_i \end{bmatrix}$. By substituting (3), (5) and (6) in (2), and some mathematical manipulations, $I_{ED}(X,Y)$ will be formulated as following:

$$\begin{aligned} &I_{ED}(X,Y) = \\ &= \sum_{i=1}^{K}\sum_{j=1}^{K} \omega_i\omega_j G(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j, C_i + C_j) \\ &+ \sum_{i=1}^{K}\sum_{j=1}^{K} \omega_i\omega_j G(\boldsymbol{\mu}_i^X - \boldsymbol{\mu}_j^X, A_i + A_j) \\ &\qquad \cdot \sum_{i=1}^{K}\sum_{j=1}^{K} \omega_i\omega_j G(\boldsymbol{\mu}_i^Y - \boldsymbol{\mu}_j^Y, B_i + B_j) \\ &-2\sum_{i=1}^{K}\sum_{j=1}^{K} \omega_i\omega_j \sum_{k=1}^{K} \omega_k G(\boldsymbol{\mu}_i - \begin{bmatrix} \boldsymbol{\mu}_j^X \\ \boldsymbol{\mu}_k^Y \end{bmatrix}, C_i + \begin{bmatrix} A_j & 0_{n\times m} \\ 0_{m\times n} & B_k \end{bmatrix}) \end{aligned} \tag{7}$$

In deriving the above expression we used the fact that the convolution of two Gaussians centered at $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ is a Gaussian centered at $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ with covariance equal to the sum of the original covariances.

2.2. Maximization of Mutual Information

Now let X be the $D-$dimensional random vector representing system inputs, and Y be the random variable for the system output. Without loss of generality we assume that the output size to be one. The goal is to find a linear transformation $g : \mathbb{R}^D \to \mathbb{R}^d$ $(d < D)$ such that the mutual information between transformed inputs and the output is maximized, or equivalently to find a $d \times D$ matrix $W^\star$ such that:

$$\begin{aligned} W^\star &= \arg\max_W \; I_{ED}(WX, Y) \\ &\text{subject to} \quad WW^T = I \end{aligned} \tag{8}$$

The constraint is set for that it makes the feasible set compact so that the existence of a mapping that maximizes the mutual information is assured. The constrained optimization problem can be changed to an unconstrained one by parameterization of the transform matrix W by Givens rotation angles. A thorough description of the method is given in [7]. Having already found the joint density of X and Y in the form of (3), it can be easily shown that the joint density of the transformed input $\tilde{X} = WX$, and the output is as following:

$$f_{\tilde{Z}}(\tilde{\boldsymbol{z}}) = \sum_{i=1}^{K} \omega_i G(\tilde{\boldsymbol{z}} - \tilde{W}\boldsymbol{\mu}_i, \tilde{W} C_i \tilde{W}^T), \tag{9}$$

with $\tilde{Z} = \begin{bmatrix} \tilde{X} \\ Y \end{bmatrix}$ and $\tilde{W} = \begin{bmatrix} W & 0_{d\times 1} \\ 0_{1\times D} & 1 \end{bmatrix}$.

Since

$$\tilde{W}\boldsymbol{\mu}_i = \begin{bmatrix} W & 0_{d\times 1} \\ 0_{1\times D} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_i^X \\ \boldsymbol{\mu}_j^Y \end{bmatrix} = \begin{bmatrix} W\boldsymbol{\mu}_i^X \\ \boldsymbol{\mu}_j^Y \end{bmatrix}$$

and

$$\tilde{W}C_i\tilde{W}^T = \begin{bmatrix} W & 0_{d\times 1} \\ 0_{1\times D} & 1 \end{bmatrix}\begin{bmatrix} A_i & D_i \\ D_i^T & B_i \end{bmatrix}\begin{bmatrix} W^T & 0_{D\times 1} \\ 0_{1\times d} & 1 \end{bmatrix} = \begin{bmatrix} WA_iW^T & WD_i \\ D_i^TW^T & B_i \end{bmatrix}$$

$I_{ED}(\tilde{X}, Y)$ can easily be calculated by replacing $\boldsymbol{\mu}_i$,$\boldsymbol{\mu}_i^X$, C_i and A_i respectively with $\tilde{W}\boldsymbol{\mu}_i$, $W\boldsymbol{\mu}_i^X$, $\tilde{W}C_i\tilde{W}^T$ and WA_iW^T wherever they appear in (7). Ending in a differentiable expression of $I_{ED}(\tilde{X}, Y)$ with respect to W, now we are able to maximize the mutual information by performing a gradient ascent method via the following iterative procedure:

$$W_{t+1} = W_t + \eta\frac{\partial I_{ED}}{\partial W} \tag{10}$$

For that, we have to compute $^{\partial I_{ED}}/_{\partial W}$. Looking at the expression of $I_{ED}(\tilde{X}, Y)$, one can observe that the first and the third terms are composed of different sums of $G(\tilde{W}\boldsymbol{\mu}, \tilde{W}C\tilde{W}^T)$ while the second term includes sums of $G(W\boldsymbol{\mu}, WCW^T)$. This implies that to compute the gradient it suffices to derive $\partial G(W\boldsymbol{\mu}, WCW^T)/\partial W$ which can be found to be:

$$\frac{\partial}{\partial W}G(W\boldsymbol{\mu}, WCW^T) = \\ -G(W\boldsymbol{\mu}, WCW^T)S^{-1}\left[(I - \boldsymbol{m}\boldsymbol{m}^TS^{-1})WC + \boldsymbol{m}\boldsymbol{\mu}^T\right] \tag{11}$$

with $\boldsymbol{m} = W\boldsymbol{\mu}$ and $S = WCW^T$. Therefore $\partial I_{ED}/\partial W$ can be computed by simply replacing the terms of form $G(W\boldsymbol{\mu}, WCW^T)$ by their derivative found above and those of form $G(\tilde{W}\boldsymbol{\mu}, \tilde{W}C\tilde{W}^T)$ by their derivative with respect to $\tilde{W}$ and eliminating the elements in the last row and the last column (as a result of the relation between W and $\tilde{W}$).

2.3. Evaluation

The estimation of the GMM model is a crucial step for our method. Given limited data samples, a too small number of mixtures will result in poor estimate while a too large one will incur overfitting. So the suitable number of mixtures should be determined based on the number of available samples. Adopting a gradient descent search for the optimization stage, our method comes to stop when the increment between two iterations is smaller than a predefined threshold. To avoid being trapped in local maxima, as an issue of gradient descent methods, we repeat searching with several random initializations for W and keep the result with the largest mutual information.

3. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of our method, we conduct experiments with both synthesized and real data sets. The advantage of using synthesized data set is our full knowledge of the system which allows us to make a fair and accurate judgement on the performance of different methods. For comparison, SIR, which gives only coarse estimate, is chosen as the baseline. We also included results for Likelihood Acquired Direction (LAD) [4], which is based on Maximum Likelihood and chosen as a representative of the state–of–the–art methods.

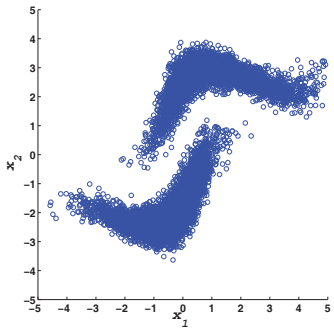

Fig. 2. Scatter plot of the inputs. 10000 point were randomly drawn from a 2-dimensional distribution.

SIR	LAD	OUR
0.4876	0.8087	0.9996

Table 1. The average estimation scores

3.1. Analysis with Synthesized Data

The input $\boldsymbol{x} = [x_1\ x_2\ x_3]^T$ is in 3-dimensional with 10000 i.i.d. samples, for which the first two dimensions are drawn from a random distribution with scatter plot in Fig. 2 while the third dimension is gaussian noise with zero mean and unit variance. We design the system output to be only relevant to the first two dimensions and the nonlinearity to be the cosine function as follows:

$$y = f(\boldsymbol{x}) = \cos(\boldsymbol{w}\boldsymbol{x}) = \cos(w_1x_1 + w_2x_2) \tag{12}$$

Expecting to get $\boldsymbol{w} = [w_1\ w_2\ 0]$ as the most relevant direction, we run SIR, LAD, our method with the transformed input in 1-dimensional. The GMM had 32 mixtures with diagonal covariance matrix. For SIR and LAD, the parameter representing the number of slices is set to 10 as we observe that the estimation results are almost the same if a larger number of slices is applied. The accuracy of each method is measured by the absolute value of the inner product of the estimated direction $\hat{\boldsymbol{w}}$ with the ground truth, i.e. $|\hat{\boldsymbol{w}}\boldsymbol{w}^T|$. A good estimate will produce a score close to 1. We run this experiment 10 times with different random $\boldsymbol{w}$, and the average of the scores of these 3 methods are presented in Table 1.

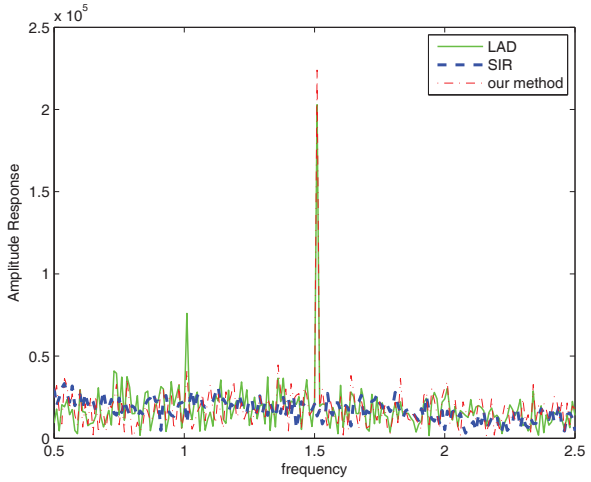

Fig. 3. Amplitude Response of the transformed recording around the modulation frequency

From Table 1, our method achieved almost perfect estimation with the third dimension of the input correctly ignored. Both LAD and SIR performed weaker in this test.

3.2. Analysis with MEG Recordings

The Magnetoencephalograph (MEG) data set [8] contains recording from 157 sensors around the brain when one subject was presented with a sinusoidally amplitude-modulated stimulus. The modulation frequency is 1.5Hz and the carrier is pure tone at 707Hz. The 2s stimulate was played 50 times, and 100s recording was collected from each sensor. The brain of this subject is a complex nonlinear system, and the sensors recorded the magnetic signals emitted from the brain' activity evolving response to the stimulus and other factors. We try to combine the recordings to obtain the "response" of the brain to the stimulus, which is modeled as a 1-dimensional waveform. Under the assumption that the response contains all the information of the stimulus, we propose to optimize the linear combination of recordings in a supervised manner. One issue should be pointed out is that the stimulus was sampled at 44.1KHz while the recordings were 500Hz. To achieve the same resolution, the stimulus was downsampled to 8KHz, and the recordings were upsampled to the same frequency without bringing distortion to the signals. The 157 recordings were first denoised with reference to the recordings from 3 additional isolated sensors to filter out background and sensor noise. We apply again SIR, LAD, and our method to reduce the 157-dimensional data to 1-dimensional. To observe the transformed recording signal in a meaningful way [8], the frequency response of this signal was obtained by Fast Fourier Transform, and the amplitude response around the modulation frequency 1.5Hz was plotted in Fig. 3.

According to [8], a peak at the modulation frequency is supposed to be seen. The response with SIR was relatively flat without any salient peak while our method and LAD successfully maintain the useful information in the transformed recording. And it is worth noting that response with our method has a higher peak at 1.5Hz (useful information) and a smaller peak at 1Hz (probably noise) compared with LAD.

4. CONCLUSION

In this paper, we propose a new method for dimension reduction in regression. Aided with GMM and Quadratic Mutual Information, we apply gradient descent method to search for the optimum linear transformation matrix by maximizing the closed-form expression of the mutual information between system output and transformed system input. Specially designed for nonlinear systems with continuous output, our method is demonstrated to be capable of estimating the linear part of the cascaded system efficiently. And promisingly, we believe it can serve as a potential tool for analyzing complex systems.

Acknowledgements. We would like to thank Jonathan Simon and Nai Ding for providing the MEG data.

5. REFERENCES

[1] J. Rapela, G. Felsen, J. Touryan, J.M. Mendel, and N.M. Grzywacz, "ePPR: a new strategy for the characterization of sensory cells from input/output data," *Network: Computation in Neural Systems*, vol. 21, no. 1-2, pp. 35–90, 2010.

[2] K. C. Li, "Sliced inverse regression for dimension reduction," *Journal of the American Statistical Association*, vol. 86, no. 414, pp. 316–327, 1991.

[3] R. D. Cook and S. Weisberg, "Discussion of "sliced inverse regression for dimension reduction "," *Journal of the American Statistical Association*, vol. 86, pp. 328–332, 1991.

[4] R. D. Cook and L. Forzani, "Likelihood-based sufficient dimension reduction," *Journal of the American Statistical Association*, vol. 104, no. 485, pp. 197–208, 2009.

[5] J. M. Leiva-Murillo and A. Artes-Rodriguez, "Maximization of mutual information for supervised linear feature extraction," *IEEE Transcations on Neural Networks*, vol. 18, no. 5, pp. 1433–1441, 2007.

[6] J.N. Kapur, *Measures of information and their applications*, Wiley, New Delhi, India, 1994.

[7] Kari Torkkola and William M. Campbell, "Mutual information in learning feature transformations," in *Proceedings of the Seventeenth International Conference on Machine Learning*, San Francisco, CA, USA, 2000, ICML '00, pp. 1015–1022.

[8] N. E. Ahmar, Y. Wang, and J. Z. Simon, "Significance tests for meg response detection," in *Proceedings of the 2nd International IEEE EMBS Conference on Neural Engineering*, Arlington, Virginia, USA, 2005.