A HYBRID SUBSPACE PROJECTION METHOD FOR SYSTEM IDENTIFICATION

Sung-Phil Kim, Yadunandana N. Rao, Deniz Erdogmus, Jose C. Principe

Computational NeuroEngineering Lab, University of Florida, Gainesville, FL 32611
[phil, yadu, deniz, principe]@cnel.ufl.edu

ABSTRACT

Principal Components Analysis (PCA) being the most optimal linear mapper in Least-Squares (LS) sense has been predominantly used in subspace-based signal processing methods. In system identification problem, optimal subspace projections must span the joint space of the input and output of the unknown system. In this scenario, subspaces determined by the principal components of the input or the desired alone do not embed key information, which lies in the joint space. In this paper, we first propose a hybrid subspace projection method that finds optimal projections in the joint space. The concepts behind this method are firmly rooted in statistical theory. We then derive adaptive learning algorithms to estimate the subspace projections. Finally, we show the superiority of the new framework in solving a system identification problem in noisy environment.

1. INTRODUCTION

In vector-space methods of signal processing, linear transformations play an important role. Linear transformations may be defined as the projection of vector-valued signals into lower-dimensional subspaces of the original data space [1]. Subspace projections play an important role in the system identification problem with noisy data as the noise blended in the signal can be reduced substantially under some constraints [2]. Principal Components Analysis (PCA), which maximally preserves the data variance, has been widely adopted as a major subspace projection method. However, in the problem of system identification, the underlying system parameters lie in the joint space of the input and desired signals. PCA, due to its very nature, cannot effectively utilize the information in joint spaces.

The subspace method (or regularization method in statistics) has been a critical issue in statistics. Many studies in statistics have formed various methods for multivariate regression to overcome the "collinearity" problem among input variables. Regularized regression methods such as Partial Least Squares (PLS), Ridge Regression (RR), and PCA are well known in statistics literature [3]. Continuum Regression (CR), introduced by Stone and Brooks [4], embraces Ordinary Least Squares (OLS), PLS, and PCA by blending their criteria. Therefore, the desirable regularization can be one of OLS, PLS, and PCA or a combination of them. The problem of system identification studied in this paper is similar to regression in statistics; hence, we may be able to utilize statistical regularization methods to design improved subspace methods.

In this paper, we propose a hybrid criterion function for subspace projection similar to CR, and develop the rules to estimate the projection matrix. We first present a gradient-based method and then improve the speed of convergence by designing a fixed-point type algorithm. We then solve the system identification problem using the new framework and proposed algorithms.

2. REVIEW OF CONTINUUM REGRESSION (CR)

In this section, we briefly summarize the criterion and the procedure of CR. Let the data be given by an input matrix $\mathbf{X}$ (n x p) and a matrix of desired responses $\mathbf{d}$ (n x m). Here, we assume m equal to 1 for simplicity. Extension to multivariate outputs can be found in [5]. Both $\mathbf{X}$ and $\mathbf{d}$ are normalized to have zero column means. In [4], the criterion is,

$$J(\mathbf{w},\alpha) = ((\mathbf{X}\mathbf{w})^T \mathbf{d})^2 ((\mathbf{X}\mathbf{w})^T \mathbf{X}\mathbf{w})^{\frac{\alpha}{1-\alpha}-1}, \|w\| = 1 \quad (1)$$

where, $\mathbf{w} \in \Re^{nx1}$, and α is a real number in the range $0 < \alpha < 1$. The special cases are $\alpha = 0$ (OLS), $\alpha = 1/2$ (PLS), and $\alpha = 1$ (PCR). If we introduce the new matrix,

$$\begin{aligned} \mathbf{R} &= \mathbf{X}^T\mathbf{X} \\ \mathbf{p} &= \mathbf{X}^T\mathbf{d} \end{aligned} \quad (2)$$

then, (1) can be rewritten as,

$$J(\mathbf{w},\alpha) = (\mathbf{w}^T\mathbf{p})^2 (\mathbf{w}^T\mathbf{R}\mathbf{w})^{\frac{\alpha}{1-\alpha}-1}, \|\mathbf{w}\| = 1 \quad (3)$$

Thus, given α, the projection weight vector $\mathbf{w}$ is constructed to maximize $J(\mathbf{w})$. After finding the first weight vector, the successive ones are computed such that weight vectors are orthogonal to each other.

3. HYBRID CRITERION FUNCTION

We build the criterion similar to CR in (3) as,

$$J(\mathbf{w},\lambda) = (\mathbf{w}^T\mathbf{p})^{2\lambda}(\mathbf{w}^T\mathbf{R}\mathbf{w})^{1-\lambda}, \|\mathbf{w}\| = 1 \quad (4)$$

where λ is a real number in the range $0 \le \lambda \le 1$. This criterion covers the continuous range between PLS ($\lambda = 1$) and PCA ($\lambda = 1$), whereas CR covers OLS, PLS and PCA. Since we are only interested in the case when subspace projection is necessary, incorporation of OLS can be omitted. To include constraint of $\|\mathbf{w}\|=1$ in the unconstrained criterion, a modified criterion $\hat{J}(\mathbf{w},\lambda)$, which is invariant to scaling of $\mathbf{w}$, can be rewritten as,

$$\hat{J}(\mathbf{w},\lambda) = \frac{(\mathbf{w}^T\mathbf{p})^{2\lambda}(\mathbf{w}^T\mathbf{R}\mathbf{w})^{1-\lambda}}{\mathbf{w}^T\mathbf{w}} \quad (5)$$

Since log is a monotonically increasing, the criterion can also be rewritten as,

$$\log(\hat{J}(\mathbf{w},\lambda)) = 2\lambda\log(\mathbf{w}^T\mathbf{p}) + (1-\lambda)\log(\mathbf{w}^T\mathbf{R}\mathbf{w}) - \log(\mathbf{w}^T\mathbf{w}) \quad (6)$$

We seek to maximize this criterion for $0 \le \lambda \le 1$. To select the best value of λ, cross-validation or other validation methods can be utilized [6].

4. LEARNING ALGORITHMS

4.1. Gradient learning algorithm

The gradient in terms of the weight vector is,

$$\frac{\partial\log(\hat{J}(\mathbf{w},\lambda))}{\partial\mathbf{w}} = \frac{2\lambda\mathbf{p}}{\mathbf{w}^T\mathbf{p}} + \frac{2(1-\lambda)\mathbf{R}\mathbf{w}}{\mathbf{w}^T\mathbf{R}\mathbf{w}} - \frac{2\mathbf{w}}{\mathbf{w}^T\mathbf{w}} \quad (7)$$

Using gradient-ascent, the weight vector can be updated as

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \eta\nabla\mathbf{l}(\hat{J}) \quad (8)$$

where $\eta > 0$ is a small learning rate. To compute the gradient, we need to know $\mathbf{R}$ and $\mathbf{p}$ that can be simply estimated from samples.

To find M such orthogonal weight vectors, where M is the desired dimension of the subspace. To find each weight vector, update rules (7) and (8) are applied on the residuals of input from the projection output in previous step (deflation procedure). Note that $\mathbf{R}$ and $\mathbf{p}$ need to be recomputed in each step. Table 1 summarizes the procedure to construct the projection weight vectors.

4.2. Fixed-point learning algorithm

The gradient method possesses a couple of defects; the speed of the convergence is slow, and the performance depends on the choice of the learning rate. Like all gradient-based methods, a finite set of step-sizes restricted by an upper bound exists for guaranteed convergence.

$\mathbf{X}$ ~ input matrix $\mathbf{d}$ ~ desired response vector M ~ dimension of projected variable (latent variable) K ~ number of iterations
Normalize $\mathbf{X}$ and $\mathbf{d}$. Given λ, For $k = 1,\ldots,K$ $\mathbf{X_0} = \mathbf{X}$ For $m = 1,\ldots,M$ Compute $\mathbf{R}$ and $\mathbf{p}$ from samples. $\nabla\mathbf{l}(\hat{J}) = \frac{2\lambda\mathbf{p}}{\mathbf{w}^T\mathbf{p}} + \frac{2(1-\lambda)\mathbf{R}\mathbf{w}}{\mathbf{w}^T\mathbf{R}\mathbf{w}} - \frac{2\mathbf{w}}{\mathbf{w}^T\mathbf{w}}$ $\mathbf{w}_m(k+1) = \mathbf{w}_m(k) + \eta\nabla\mathbf{l}(\hat{J},k)$ $\mathbf{w}_m(k+1) = \frac{\mathbf{w}_m(k+1)}{\|\mathbf{w}_m(k+1)\|}$ $\mathbf{y}_m = \mathbf{X}_{m-1}\mathbf{w}_m(k+1)$ $\mathbf{X}_m = \mathbf{X}_{m-1} - \mathbf{y}_m\frac{\mathbf{y}_m^T\mathbf{X}_{m-1}}{\mathbf{y}_m^T\mathbf{y}_m}$ end end

Table 1. Gradient hybrid subspace learning algorithm.

However, this step-size upper bound is data dependent which makes it unwieldy to utilize in any practical applications. Recently fixed-point algorithms have been proposed for PCA that were shown to converge faster than gradient methods [7]. Motivated by the effectiveness of the fixed-point PCA rules, we derived the fixed-point version of the hybrid subspace learning algorithm. The stationary point of (7) is given by equating the gradient to zero.

Assuming that $\|\mathbf{w}\|=1$, we can rearrange the terms as,

$$\mathbf{w} = \left[\frac{\lambda\mathbf{p}}{\mathbf{w}^T\mathbf{p}} + \frac{(1-\lambda)\mathbf{R}\mathbf{w}}{\mathbf{w}^T\mathbf{R}\mathbf{w}}\right] \quad (9)$$

Let the weight vector $\mathbf{w}(k)$ be the estimation of the first projection direction at iteration k. Then the estimate of the weight vector at iteration index k+1 is,

$$\mathbf{w}(k+1) = \left[\frac{\lambda\mathbf{p}}{\mathbf{w}(k)^T\mathbf{p}} + \frac{(1-\lambda)\mathbf{R}\mathbf{w}(k)}{\mathbf{w}(k)^T\mathbf{R}\mathbf{w}(k)}\right] \quad (10)$$

where, $\mathbf{w}(k)^T\mathbf{w}(k) = 1$. It can be shown using the principles outlined in [7] and [8] that the algorithm in (10) will enter a limit cycle (near convergence) resulting in the oscillation of $\mathbf{w}$ between two vectors. To remove the oscillation behavior of weight vector when the convergence is reached, we can balance the previous value of the weights with the new correction as shown in [2],

$$\mathbf{w}(k+1) = (1-T)\mathbf{w}(k) + T\left[\frac{\lambda\mathbf{p}}{\mathbf{w}(k)^T\mathbf{p}} + \frac{(1-\lambda)\mathbf{R}\mathbf{w}(k)}{\mathbf{w}(k)^T\mathbf{R}\mathbf{w}(k)}\right] \quad (11)$$

where $0 < T < 1$. The convergence rate is affected by T, which produces a tradeoff between the convergence speed and the accuracy. The overall procedure for finding subsequent projection weight vectors is the same as the gradient algorithm depicted in table 1. Only the equation (8) in the table is changed with (11).

We now investigate the convergence characteristics of the fixed-point update rule. The ordinary differential equation equivalent to (11) is,

$$\frac{\partial \mathbf{w}(t)}{\partial t} \approx \frac{\mathbf{w}(k+1)-\mathbf{w}(k)}{T}$$
$$= \frac{\lambda \mathbf{p}}{\mathbf{w}(k)^T \mathbf{p}} + \frac{(1-\lambda)\mathbf{R}\mathbf{w}(k)}{\mathbf{w}(k)^T \mathbf{R}\mathbf{w}(k)} - \mathbf{w}(k) \quad (12)$$

Then, we can state the following theorem;

Theorem 1: The norm of the weight vector, $\|\mathbf{w}(t)\|^2$ converges to 1 as $t \to \infty$.

Proof: Multiplying from the left by $\mathbf{w}^T(t)$ on both sides of (12), we get,

$$\mathbf{w}^T(t)\frac{\partial \mathbf{w}(t)}{\partial t} = [1 - \mathbf{w}^T(t)\mathbf{w}(t)] = [1 - \|\mathbf{w}(t)\|^2] \quad (13)$$

From the above equation, it is easy to see that,

$$\frac{\partial \|\mathbf{w}(t)\|^2}{\partial t} = 2\mathbf{w}^T(t)\frac{\partial \mathbf{w}(t)}{\partial t} = 2[1 - \|\mathbf{w}(t)\|^2] \quad (14)$$

The above ODE can be easily solved as,

$$\|\mathbf{w}(t)\|^2 = 1 - (1 - \|\mathbf{w}(0)\|^2)e^{-2t} \quad (15)$$

where $\mathbf{w}(0)$ denotes the initial weight vector. As $t \to \infty$, the norm converges to unity. This completes the proof.

Theorem 2: The fixed-point update equation converges to a stable stationary point in the Lyapunov stability sense if $1 < \|\mathbf{w}(0)\|^2 < \infty$.

Proof: Suppose we choose a positive definite Lyapunov function as $L(t) = \mathbf{w}(t)^T\mathbf{w}(t)$. Then, the $\partial L(t)/\partial t$ becomes,

$$\frac{\partial L(t)}{\partial t} = 2\mathbf{w}(t)^T \frac{\partial \mathbf{w}(t)}{\partial t}$$
$$= 2(1 - \|\mathbf{w}(t)\|^2) \quad (13)$$

$\partial L(t)/\partial t \le 0$ for all t only if $\|\mathbf{w}(t)\|^2 \to 1$. But, from theorem 1, we know that the norm is either a monotonically increasing or decreasing function with unity being the stationary point. Therefore, if $1 < \|\mathbf{w}(0)\|^2 < \infty$, then, $\partial L(t)/\partial t \le 0$ for all t, and eventually when $\|\mathbf{w}(t)\|^2 \to 1$ as $t \to \infty$, L(t) becomes zero.

The complete proof of the convergence of the fixed-point update equation will be present in another work.

5. EXPERIMENTS

We will present a system identification problem in the presence of noise. Traditional MSE-based techniques fail because of the presence of noise in the data. Subspace algorithms on the other hand can find optimal projections that can suppress the noise leading to more accurate identification. A continuous signal comprising of three sinusoidal components was generated as $x(t) = c_1\sin(2\pi f_1 t) + c_2\sin(2\pi f_2 t) + c_3\sin(2\pi f_3 t)$, where f_1 is 10Hz, f_2 is 50Hz, and f_3 is 100Hz, respectively. The sampling rate was 1kHz. The coefficients c_1, c_2 and c_3 can be set arbitrary (e.g. $c_1 = c_2 = c_3 = 1$ in our experiment). This signal was filtered through a real-valued FIR filter with 10-taps (target filter) to generate the desired response signal. Then white Gaussian noise was added to $x(t)$ to produce the noisy data. One thousand 50-dimensional vectors were derived out of this using a tap-delay line to create an input matrix with dimension of 1000×50. The subspace dimension was chosen as 6 (as there were 3 sinusoids). Then, the reconstruction matrix, which is equal to the transpose of a subspace matrix, linearly combined the subspace signal to form the reconstructed signal. Note that the reconstructed signal contains less noise due to the subspace projection. The reconstructed signal was then used as an input to a 10-tap adaptive filter to estimate the target filter. The overall architecture is depicted in Fig. 1. In simulation, we first trained a subspace projection matrix

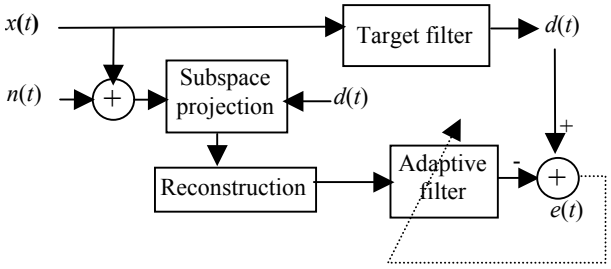

Figure 1. The structure of the system identification example with subspace projection and reconstruction. The adaptive filter with reconstructed signal is the estimation of the target filter.

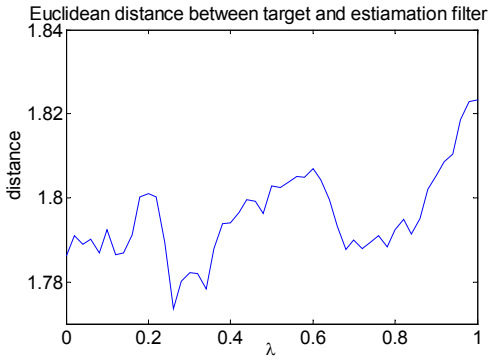

Figure 2. Euclidean distance between a target filter's coefficient vector and an estimated filter's over $0 \le \lambda \le 1$.

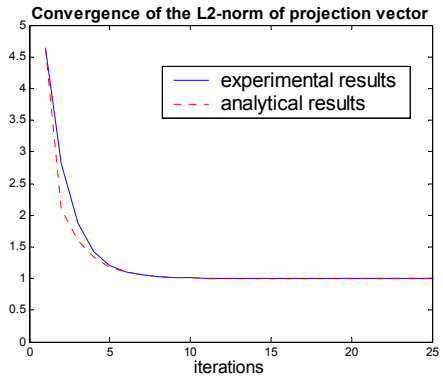

Figure 3. The convergence of the L2 norm of subspace projection weight vector.

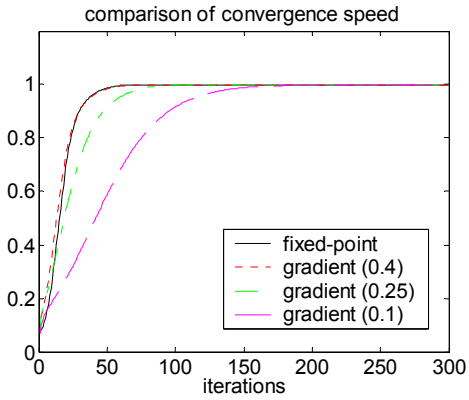

Figure 4. The comparison of the convergence speed between the fixed-point with T as 0.9, and the gradient rule with learning rate as 0.1, 0.25, and 0.4, respectively.

using the two proposed algorithms for 50 epochs, and then trained an adaptive filter by LMS.

First, the Euclidean distances between the target filter evaluated as shown in Fig. 2. The results were obtained from average of 50 Monte Carlo simulations. The best estimation result was observed when λ was approximately 0.26. This indicates that we can estimate the target filter more accurately using hybrid subspace projection than PCA. Next, we verify that L_2 norm of the weight vector converges to 1. The experimental results of $\|\mathbf{w}(t)\|^2$ for $T = 0.3$, as t increases, are plotted in Fig. 3, and compared with analytically computed values from (15). As a result, two convergence curves from experimental and analytical computations are very similar to each other. Finally, the convergence speed between the fixed-point and the gradient rule were compared when $\lambda = 0$ as illustrated in Fig. 4. The projection weight vector converges to the eigenvector of the input correlation matrix in this case. The absolute value of $\cos\theta$, where θ is the angle between the actual eigenvector and the weight vector, was computed over iterations. It shows that the fixed-point rule converges faster when a learning rate was less than 0.4.

According to the experimental results, it can be stated that there exists a better linear subspace projection than PCA if we exploit the desired response.

6. DISCUSSIONS

We have presented a hybrid subspace projection framework that effectively uses information in the joint space of a pair of signals. The proposed cost function included PLS and PCA as special cases. We then proposed gradient as well as fixed-point type algorithm to maximize the hybrid cost function. In order to verify the power of this hybrid framework, we performed a system identification experiment in noisy environment. Our results showed that the new approach has better noise rejection capability when compared with the traditional PCA based subspace methods.

The proposed algorithm may be useful for the pattern recognition or signal detection problem. In that case, we might need to consider how to manipulate the qualitative desired response (not quantitative) to compute the cross-validation vector, **p** in (5). Future studies will cover this problem.

7. REFERENCES

[1] E. Oja, *Subspace Methods of Pattern Recognition,* Research Studies Press, England, 1983.

[2] Y.N. Rao, "Adaptive Eigendecomposition Algorithms for Time Series Segmentation," MS Thesis, University of Florida, 2000.

[3] R. Sundberg, "Shrinkage regression," *Encyclopedia of Environmetrics*, vol. 4, pp. 1994-1998, 2002.

[4] M. Stone, and R.J. Brooks, "Continuum regression: Cross-validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression (with discussion)," *J. Royal Statist. Soc. Ser. B*, 52, pp. 237-269, 1990.

[5] R.J. Brooks, and M. Stone, "Joint Continuum Regression for Multiple Predictands," *J. Amer. Statist. Assoc.*, 89, pp. 1374-1377, 1994.

[6] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, NJ, 1999.

[7] Y.N. Rao, J.C. Principe, "A Fast, On-line Algorithm for PCA and its Convergence Characteristics," *Proc. IEEE Workshop on Neural Networks for Signal Processing X*, pp. 299-308, 2000.

[8] Y.N. Rao, J.C. Principe, "Robust On-line Principal Component Analysis Based on a Fixed-Point Approach", *Proc. ICASSP*, vol. 1, pp. 981-984, 2002.