Comparison between MSE and MEE Based Component Extraction Approaches to Process Monitoring and Fault Diagnosis

Zhenhua Guo[‡,1] and Hong Wang[$]

[‡]School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, Hubei, PRC
[$] Department of Electrical Engineering and Electronics, University of Manchester, UK

Abstract Component extraction is a technique for extracting the latent components that underlie the observation of a set of variables. In the paper both classical Principal component analysis (PCA) and autoassociative principal component neural network (PCNN) methods with minimum mean square error (MSE) criterion are compared with the corresponding extended component extraction methods with Minimum error entropy (MEE) criterion in theory. A Parzen window estimator based approximative computation method for entropy is provided, and the equivalence between MSE and MEE criteria is also analyzed.

Key words: principal component analysis; minimum square error; minimum error entropy; fault diagnosis, neural network

1. INTRODUCTION

Various multivariate statistical techniques have been employed for revealing the statistical characteristics of processes, including PCA (Jolliffe, 2002), principal component neural networks (PCNN) (Diamantaras and Kung, 1996), Fisher discriminant analysis (He et al., 2004) and independent component analysis (ICA) (Hyvärinen et al, 2001) et al. Multivariate statistical process control (MSPC) techniques based on PCA and its extensions are the most widely applied in process monitoring and fault diagnosis (MacGregor and Koutodi, 1995, Nomikos and MacGregor, 1995). PCA is an effective statistical technique for data compression and component extraction with the optimal performance in the sense of minimum mean square error (MSE), and the extracted principal components are the combinations of variables that describe major trends in a data set. Low dimensional features enable the process control and monitoring to use the less computational resources and memory without compromising the accuracy. Since the Gaussian probability density function (PDF) is determined only by its first- and second-order statistics. Under these linearity and Gaussianity assumptions, further supported by the central limit theorem, MSE criterion would be able to extract all possible information from a signal whose statistics are solely defined by its mean and variance. Therefore, MSE is a very suitable index for Gaussian distribution systems, but may be inappropriate for the non-Gaussian distribution systems.

However, achievable performance of PCA is limited mainly due to the assumption that the multivariate data are normally distributed. To further improve the analytical performance for non-Gaussian dataset, several MSPC methods based on high order statistic and neural network have been proposed (Diamantaras and Kung, 1996; Girolami, 1999), including nonlinear principal component neural network and independent component analysis (ICA). Entropy is an alternative general criterion for the measure of the uncertainty in non-Gaussian system, and can be used as an alternative criterion to the MSE criterion (Taylor and Plumbley, 1993, Wang, 2002).

2. PCA OF MINIMUM SQUARE ERROR CRITERIA

Consider a continuous random vector $\mathbf{x} = [x_1\ x_2 \cdots x_n]^T \in R^n$ with the mean $E\{\mathbf{x}\} = 0$ and covariance matrix $\mathbf{R}_\mathbf{x} = E\left\{\mathbf{x}\mathbf{x}^\mathbf{T}\right\} \in R^{n\times n}$.. If the eigenvalues of the covariance matrix $\mathbf{R}_\mathbf{x}$ are ascendingly ordered as $\lambda_1\lambda_2 \cdots \lambda_n > 0$ and the corresponding eigenvectors are $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n$, the transform $y_i = \mathbf{e}_i^T\mathbf{x}$ is called i th principal component (PC) or score and the eigenvector $\mathbf{e}_i$ is called principal eigenvector or loading (Diamantaras and Kung, 1996, Hyvärinen et al., 2001). For any i th and j th PCs, they have the variance $\mathrm{var}(y_i) = \lambda_i$ and the covariance $\mathrm{cov}(y_i, y_j) = 0\,(i \neq j)$. Therefore, the first principal component has the largest variance and the principal components are mutually uncorrelated. PCA can also be presented from discrete sample data. For a discrete observation matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in R^{N\times n}$ that consists of N

observations with n variables, PCA decomposes $\mathbf{X}$ into the sum of the outer product of n pairs of vectors as follows

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \cdots + \mathbf{t}_n\mathbf{p}_n^T = \mathbf{T}\mathbf{P}^T \quad (1)$$

Let the eigenvalues of covariance matrix $\mathbf{R}_x = E\{\mathbf{X}\mathbf{X}^T\} \in R^{n\times n}$ be arranged in decreasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and let their corresponding normalized eigenvectors be denoted by $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_m$. According to the Rayleigh-Ritz theorem (Diamantaras and Kung,1996), these eigenvectors and eigenvalues are equal to the loadings and the variances of principal components (PCs) in Equation (1).. That is

$$e_i = p_i; \lambda_i = \mathrm{var}(t_i) \quad (2)$$

Different criteria had been proposed in the past for the selection of the optimal number of PCs.. In this paper the cumulative percent variance (CPV) (Valle et al., 1999) based on the cumulative variance contribution rate of the first several PCs is used to choose the number of PCs, which is defined as follows

$$CPV = \frac{\sum_{i=1}^{k} \mathrm{var}(t_i)}{\sum_{j=1}^{n} \mathrm{var}(t_j)} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{n} \lambda_j} \quad (3)$$

where k is the number of PCs reserved. When PCA is performed on a data matrix, it is often found that only the first k PCs are associated with system variation in the data and the remaining $m-k$ PCs are associated with the noise. With these reserved PCs, which are descriptive of system variation, the following PCA model can be formulated.

$$\hat{\mathbf{X}} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \cdots + \mathbf{t}_k\mathbf{p}_k^T = \mathbf{T}_k\mathbf{P}_k^T \quad (4)$$

$$\mathbf{E} = \mathbf{t}_{k+1}\mathbf{p}_{k+1}^T + \mathbf{t}_{k+2}\mathbf{p}_{k+2}^T + \cdots + \mathbf{t}_n\mathbf{p}_n^T \quad (5)$$

where $\hat{\mathbf{X}}$ is the estimator of $\mathbf{X}$, and $\mathbf{E}$ is the corresponding residual error matrix. Multivariate control charts are used to detect shifts in the means of variables or the relationship (covariance) between several related parameters. The most extensively applied multivariate control charts are squared prediction error (SPE) chart, Hotelling's T^2 control chart, contribution charts and loading chart (MacGregor and Koutodi, 1995, Nomikos and MacGregor, 1995). SPE also called Q statistic and Hotelling's T^2 statistic are often used to detect shifts in the process. Q statistic offers a way to test if the process data has shifted outside the normal operating space, while T^2 statistic provides an indication of unusual variance within the normal subspace. For a new observation $\mathbf{x}_i$, Q_i and T_i^2 statistics are defined as follows.

$$Q_i = \mathbf{x}_i(\mathbf{I} - \mathbf{P}_k\mathbf{P}_k^T)\mathbf{x}_i^T \quad (6)$$

$$T_i^2 = \mathbf{x}_i\mathbf{P}_k\Lambda_k\mathbf{P}_k^T\mathbf{x}_i^T \quad (7)$$

where $\Lambda_k = \mathrm{diag}\{\lambda_1, \cdots, \lambda_k\}$. The upper control limits (UCLs) for SPE and T^2 with α confidence can be calculated respectively as follows(MacGregor and Koutodi, 1995, Nomikos and MacGregor, 1995, Jackson and Mudholkar, 1979)..

$$Q_\alpha = \theta_1\left[\frac{C_\alpha\sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0(h_0-1)}{\theta_1^2}\right]^{\frac{1}{h_0}} \quad (8)$$

$$T^2_{m,N,\alpha} = \frac{m(N-1)}{(N-m)}F_{m,N-1,\alpha} \quad (9)$$

$$\theta_i = \sum_{j=m+1}^{n}\lambda_j^i \quad (i = 1,2,3) \quad (10)$$

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2} \quad (11)$$

n and N are the number of variables or PCs and the number of samples respectively. m is the number of the reserved PCs. λ_i is the eigenvalue of covariance matrix $\mathbf{R}_x = E\{\mathbf{X}\mathbf{X}^T\}$. C_α is the upper α th quantile of standard normal distribution. $F_{m,N-1,\alpha}$ represents the F distribution with freedom parameters m and $N-1$. After performing PCA on a random vector $\mathbf{x} \in R^n$ with the mean $E\{\mathbf{x}\} = 0$ and covariance matrix $\mathbf{R}_\mathbf{x} \in R^{n\times n}$, and reserving the first k principal components and the corresponding orthogonal eigenvectors denoted by $\mathbf{W} \in R^{k\times n}$, the following principal component vector $\mathbf{y}$ can be obtained.

$$\mathbf{y} = \mathbf{W}\mathbf{x} \quad (12)$$

Denote the k-dimensional subspace spanned by $\mathbf{W}$ as $\mathrm{span}(\mathbf{W})$. The projection of $\mathbf{x}$ onto subspace $\mathrm{span}(\mathbf{W})$, which is the reconstruction of $\mathbf{x}$ from $\mathbf{y}$, can be expressed as follows (Diamantaras and Kung, 1996, Hyvärinen, et al., 2001).

$$\hat{\mathbf{x}} = \mathbf{W}^T\mathbf{y} \quad (13)$$

PCA seeks to minimize the mean square reconstruction error (Diamantara and Kung, 1996)..

$$J_{MSE} = E\left\{\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|^2\right\} = E\left\{\mathrm{tr}\left((\mathbf{x}-\hat{\mathbf{x}})(\mathbf{x}-\hat{\mathbf{x}})^T\right)\right\} \quad (14)$$

where $\mathrm{tr}(\cdot)$ is the trace operator. According to the fact that $\mathrm{tr}(\mathbf{A}\mathbf{B}) = \mathrm{tr}(\mathbf{B}\mathbf{A})$, $\mathbf{A} \in R^{n\times k}$ and $\mathbf{B} \in R^{k\times n}$, Equation (15) can be further written as

$J_{mse} = \mathrm{tr}$

$$J_{mse} = \mathrm{tr}(\mathbf{R_x}) - \mathrm{tr}(\mathbf{W R_x W^T}) \quad (15)$$

$$\mathrm{tr}(\mathbf{W}\,\mathbf{R}_{\mathbf{x}}\mathbf{W}^{T}) = E\left\{\mathrm{tr}(\mathbf{y}\mathbf{y}^{T})\right\} = E\left\{\mathrm{tr}(\hat{\mathbf{x}}\hat{\mathbf{x}}^{T})\right\} \quad (16)$$

3 PCA NN OF MINIMUM SQUARE ERROR

Neural network (NN) provides a novel way for parallel online computation of PCA, and a number of exciting advantages have been obtained. An auto-associative neural network has the feature of the same number of neurons in the input and output layers and the less number of neurons in the hidden layers, is trained using the input vector itself as the desired output. This training leads to organize a compression or encoding network between the input layer and the hidden layer, and a decoding network between the hidden layer and the output layer. It turns out that there is indeed a close relationship between PCA and networks of this type.

Denote the input of network as $\mathbf{X} \in R^{n\times N}$ and the output of hidden layer as $\mathbf{H} \in R^{m\times N}$. The output of networks, which is the reconstruction of $\mathbf{X}$, can be write as

$$\hat{\mathbf{X}} = \overline{\mathbf{W}}^{T}\mathbf{H} + \overline{\theta}\mathbf{u}^{T} \quad (17)$$

$$\mathbf{H} = \mathrm{F}\left(\underline{\mathbf{W}}^{T}\mathbf{X} + \underline{\theta}\mathbf{u}^{T}\right) \quad (18)$$

where $\mathbf{u} = [1,1,\cdots,1]^{T} \in R^{N\times 1}$ and $\mathrm{F}(\cdot)$ is the nonlinear activation function operating on each element of the vector. $\underline{\mathbf{W}} \in R^{n\times m}$ and $\overline{\mathbf{W}} \in R^{m\times n}$ are the weight matrices of the hidden and the output layers respectively. $\underline{\theta} \in R^{m\times 1}$ and $\overline{\theta} \in R^{n\times 1}$ are the bias vectors of the hidden layer and the output layer respectively. Using the square error cost, the performance index becomes (Diamantara and Kung, 1996)

$$J = \frac{1}{2N}\left\|\mathbf{X} - \hat{\mathbf{X}}\right\|^{2} = \frac{1}{2N}\left\|\mathbf{X} - \overline{\mathbf{W}}^{T}\mathbf{H} - \overline{\theta}\mathbf{u}^{T}\right\|^{2} = \frac{1}{2N}\sum_{i=1}^{N}\left\|\mathbf{x}_{i} - \overline{\mathbf{W}}^{T}\mathbf{h}_{i} - \overline{\theta}\right\|^{2} \quad (19)$$

The learning algorithm for the network is to select the optimal $\underline{\mathbf{W}}$, $\overline{\mathbf{W}}$, $\underline{\theta}$ and $\overline{\theta}$, so that the square error cost is minimized. Setting

$$\frac{\partial J}{\partial \overline{\theta}} = -\frac{1}{N}\sum_{k=1}^{N}(\mathbf{x}_{i} - \overline{\mathbf{W}}^{T}\mathbf{h}_{i} - \overline{\theta}) = 0 \quad (20)$$

The optical $\overline{\theta}$ can be obtained

$$\overline{\theta} = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{x}_{i} - \overline{\mathbf{W}}^{T}\mathbf{h}_{i}) = \frac{1}{N}(\mathbf{X} - \overline{\mathbf{W}}^{T}\mathbf{H})\mathbf{u} \quad (21)$$

which submitted into the cost function yields

$$J = \frac{1}{2N}\left\|\mathbf{X}(\mathbf{I} - \mathbf{u}\mathbf{u}^{T}/N) - \overline{\mathbf{W}}^{T}\mathbf{H}(\mathbf{I} - \mathbf{u}\mathbf{u}^{T}/N)\right\| = \frac{1}{2N}\left\|\mathbf{X}' - \overline{\mathbf{W}}^{T}\mathbf{H}'\right\| \quad (22)$$

$\mathbf{X}' = \mathbf{X}(\mathbf{I} - \mathbf{u}\mathbf{u}^{T}/N)$, and $\mathbf{H}' = \mathbf{H}(\mathbf{I} - \mathbf{u}\mathbf{u}^{T}/N)$.

Since $\overline{\mathbf{W}} \in R^{m\times n}$ and $rank(\overline{\mathbf{W}}) \le m$, $rank(\overline{\mathbf{W}}^{T}\mathbf{H}') \le m$, the problem of minimization of J is equivalent to the problem of the best rank-p approximation of X'. Using singular value decomposition (SVD), X' can be decomposed as

$$\mathbf{X}' = \mathbf{U}_{n}\Sigma_{n}\mathbf{V}_{n}^{T} \quad (23)$$

where $\Sigma_{n} = diag[\sigma_{1},\sigma_{2},\cdots,\sigma_{n},0,\cdots,0] \in R^{n\times N}$, and $\sigma_{1} \ge \sigma_{2} \ge \cdots \ge \sigma_{n}$. σ_{i} is the singular value of $\mathbf{X}'$. $\mathbf{U}_{n} \in R^{n\times n}$ and $\mathbf{V}_{n} \in R^{N\times N}$ are the matrixes containing the corresponding singular vectors. According to the results in (Achlioptas and McSherry, 2001), if $\mathbf{U}_{m} \in R^{n\times m}$ and $\mathbf{V}_{m} \in R^{N\times m}$ are formed by the first m columns from $\mathbf{U}_{n}$ and $\mathbf{V}_{n}$ respectively, and the diagonal matrix $\Sigma_{m} = diag[\sigma_{1},\sigma_{2},\cdots,\sigma_{m}]$ is the partial matrix containing the m largest singular values, the following expression is the best rank-p approximation of X'.

$$\overline{\mathbf{W}}^{T}\mathbf{H}' = \mathbf{U}_{m}\Sigma_{m}\mathbf{V}_{m}^{T} \quad (24)$$

$$\overline{\mathbf{W}}^{T} = \mathbf{U}_{m}\Lambda; \quad \mathbf{H}' = \Lambda^{-1}\Sigma_{m}\mathbf{V}_{m}^{T} \quad (25)$$

where $\Lambda = R^{m\times m}$ is any nonsingular matrix..

4. PCA OF MINIMUM ERROR ENTROPY

Entropy is a probabilistic measure of uncertainty (Girolami, 1999). Let $\mathbf{x} \in R^{n}$ be a continuous random vector with a probability density function $p_{\mathbf{x}}(\mathbf{x})$. The differential entropy of $\mathbf{x}$ is defined as

$$H(\mathbf{x}) = -\int_{\mathbf{R}^{n}} p(\mathbf{x})\log p(\mathbf{x})d\mathbf{x} \quad (26)$$

If $\mathbf{x}$ an n-dimensional Gaussian distribution vector with zero mean and covariance matrix $\mathbf{R}_{\mathbf{x}}$.

$$H(\mathbf{x}) = \frac{1}{2}\log(2\pi e)^{n}\left|\mathbf{R}_{\mathbf{x}}\right| \quad (27)$$

where $\left|\mathbf{R}_{\mathbf{x}}\right|$ denotes the determinant of $\mathbf{R}_{\mathbf{x}}$. For two continuous random vectors $\mathbf{x} \in R^{n}$ and $\mathbf{y} \in R^{m}$, the conditional entropy $H(\mathbf{x}|\mathbf{y})$ is defined as

$$H(\mathbf{x}|\mathbf{y}) = -\int_{R^{n}}\int_{R^{m}} p(\mathbf{x},\mathbf{y})\log p(\mathbf{x}|\mathbf{y})d\mathbf{x}d\mathbf{y} \quad (28)$$

where $p(\mathbf{x}|\mathbf{y})$ is the conditional probability density of $\mathbf{x}$ given $\mathbf{y}$. The mutual information between two continuous random vectors $\mathbf{x}$ and $\mathbf{y}$ with joint

probability density $p(\mathbf{x},\mathbf{y})$ is defined as

$$I(\mathbf{x};\mathbf{y}) = \int_{R^n}\int_{R^m} p(\mathbf{x},\mathbf{y}) \log \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})\,p(\mathbf{y})} d\mathbf{x}d\mathbf{y} \qquad (29)$$

where $p(\mathbf{x})$ and $p(\mathbf{y})$ are the marginal probability density for $\mathbf{x}$ and $\mathbf{y}$ respectively. For a random vector $\mathbf{x} = [x_1\ x_2 \cdots x_n]^T \in R^n$, the mutual information can be wrote as

$$I(x_1, x_2, \cdots x_n) = \sum_{i=1}^{n} H(x_i) - H(\mathbf{x}) \qquad (30)$$

From the definitions of entropy, conditional entropy and mutual information, the following equation can be got (Girolami, 1999).

$$I(\mathbf{x};\mathbf{y}) = H(\mathbf{x}) - H(\mathbf{x}\,|\,\mathbf{y}) = I(\mathbf{y};\mathbf{x}) \qquad (31)$$

Consider a random vector $\mathbf{x}$ with sample dataset $\mathbf{A}$, the Parzen-window estimator can be written as follows (Parzen, 1962).

$$\hat{p}(\mathbf{x}, A) = \frac{1}{N_A}\sum_{\mathbf{x}_A \in A} \kappa(\mathbf{x} - \mathbf{x}_A) = E_A[\kappa(\mathbf{x} - \mathbf{x}_A)] \qquad (32)$$

where $\mathbf{x}_\mathrm{A}$ is a sample in $\mathbf{A}$, N_A is the number of samples used to estimate the density and $E_\mathrm{A}[\cdot]$ is the mean values evaluated from the samples. $\kappa(\cdot)$ is the kernel function, which usually uses Gaussian probability density function. According to the PDF estimator in Equation (32), the entropy in Equation (28) can be rewritten as follows.

$$H(\mathbf{x}) \approx -E\left[\log \hat{p}(\mathbf{x}, A)\right] \qquad (33)$$

Substituting the integral $E\left[\log \hat{p}(\mathbf{x}, A)\right]$ with the mean value of the samples, the approximative entropy can be further expressed as

$$H(\mathbf{x}) \approx -E_B\left[\log \hat{p}(\mathbf{x}, A)\right] = -\frac{1}{N_B}\sum_{\mathbf{x}_B \in B} \log\left\{\frac{1}{N_A}\sum_{\mathbf{x}_A \in A}\kappa(\mathbf{x}_B - \mathbf{x}_A)\right\} \qquad (34)$$

where N_A is the number of samples used to estimate density, while N_B is the number of samples used to approximate entropy.. Let us consider n-dimensional random vector $\mathbf{x} \in R^n$ with zero mean and covariance matrix $\mathbf{R}_\mathbf{x} = E\{\mathbf{x}\mathbf{x}^T\}$. From the PCA model given by Equation (4) the following principal component vector can be obtained.

$$\mathbf{y} = \mathbf{W}^T\mathbf{x} \qquad (35)$$

where $\mathbf{W} \in R^{n\times k}$ is an orthogonal matrix whose i th column is the i th eigenvector of $\mathbf{R}_\mathbf{x}$. Under the n-dimensional Gaussian distribution assumption for $\mathbf{x}$, $\mathbf{y}$ also is k-dimensional Gaussian distribution with covariance matrix $\mathbf{R}_\mathbf{x} = \mathbf{W}^T\mathbf{R}_\mathbf{x}\mathbf{W}$. The entropy of $\mathbf{y}$ can be calculated via Equation (27). Apparently, $H(\mathbf{y})$ is proportional to the determinant of covariance matrix $\mathbf{R}_\mathbf{y}$. Using the results in (Jolliffe, 2002), it can be seen that determinant of $\mathbf{R}_\mathbf{y}$ is the maximum. Therefore, minimizing the mean square error or maximizing the PCs variance for Gaussian system by PCA is equivalent to the minimization of the error entropy or the maximization of the model entropy. Since entropy has more general meaning than that of variance for arbitrary random variables, it can be used to measure the uncertainty of random variables and form a design criteria for general stochastic system subjected to arbitrary distribution (Wang, 2002). Based on this idea a modified PCA technique with minimum error entropy (MEE-PCA) has been proposed (Guo and Wang, 2004) MEE-PCA performs a conventional PCA on the data and chooses the number of PCs reserved via CPV, and then optimizes the corresponding loading vector by a genetic algorithm with the minimum errors entropy.

To be more specific, after PCA with the first k reserved PCs, the PCA model as Equation (4) can be obtained. In order to formulate the best model with minimum error entropy, the reserved loading eigenvectors in $\mathbf{P}_k$ shall be optimized via a genetic algorithm. For the convenience of chromosome coding in genetic algorithm, the reserved loading eigenvectors in $\mathbf{P}_k$ are rearranged as a value vector G for the optimization.

$$\mathbf{P}_k = \begin{bmatrix}\mathbf{p}_1\\ \mathbf{p}_2\\ \vdots\\ \mathbf{p}_k\end{bmatrix} \Longleftrightarrow \mathbf{G} = \begin{bmatrix} g_1\\ g_2\\ \vdots\\ g_{n\times k}\end{bmatrix} \qquad (36)$$

where there is a mapping of elements between $\mathbf{p}_i$ and $\mathbf{G}$ as follows.

$$\mathbf{p}_i = \begin{bmatrix} p_{1i}\\ p_{2i}\\ \vdots\\ p_{ni}\end{bmatrix} \Longleftrightarrow \begin{bmatrix} g_{(i-1)\times n+1}\\ g_{(i-1)\times n+2}\\ \cdots\\ g_{i\times n}\end{bmatrix} \qquad (37)$$

. In order to further optimize the error entropy, an optimization model with a random vector pertinent to the loading eigenvectors in Equation (36) shall be built. This random vector can also be uniquely converted back to a random loading eigenvector matrix. Denoting the random vector as $\mathbf{U}$ and its relevant loading eigenvector matrix as $\mathbf{L}$, the following

optimization model can be formulated.

$$\mathbf{U}=\begin{bmatrix} u_1 \\ u_2 \\ \cdots \\ u_{n\times k} \end{bmatrix} \Longleftrightarrow \mathbf{L}=\begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1k} \\ l_{21} & l_{22} & \cdots & l_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ l_{n1} & l_{n2} & \cdots & l_{nk} \end{bmatrix} \quad (38)$$

In order to optimize the random variable in $\mathbf{U}$ via the genetic algorithm, the domain of the random variable shall be decided. For the reason that the random variables are used to determinate the directions of loading eigenvectors, their values are proportional to each other. The domain for the random vector $\mathbf{U}$ in Equation (38) can be set to $[0.5\mathbf{G},\ 1.5\mathbf{G}]$, i.e. $u_i \in [0.5g_i, 1.5g_i]$ $(1 \le i \le (n\times k))$. In the meantime the loading eigenvectors in $\mathbf{L}$ shall also be subjected to

$$\begin{aligned} l_i^T l_j &= 0 \quad (i \ne j;\ 1 \le i, j \le (n\times k)) \\ l_i^T l_j &= 1 \quad (i = j;\ 1 \le i, j \le (n\times k)) \end{aligned} \quad (39)$$

The more detail introduction and implementation of this MEE-PCA method can be refereed from (Guo and Wang, 2004).

5. A NN BASED PCA WITH MINIMUM ERROR ENTROPY

In this section an auto-associative neural network with a topology similar to autoassociative MSE-based PCNN model and the corresponding learning algorithm are presented.

Let $\mathbf{x} \in R^n$ and $\hat{\mathbf{x}} \in R^n$ be the input and the output of the neural network, and $\mathbf{h} \in R^m$ $(m \le n)$ and $\mathrm{F}(\square)$ be the output and the nonlinear activation function of hidden layer respectively. The following transitional equations can be obtained.

$$\mathbf{h} = \mathrm{F}(\underline{\mathbf{W}}^T\mathbf{x} + \underline{\theta}) \quad (40)$$

$$\hat{\mathbf{x}} = \overline{\mathbf{W}}^T\mathbf{h} + \overline{\theta} \quad (41)$$

$$\varepsilon = \hat{\mathbf{x}} - \mathbf{x} = \overline{\mathbf{W}}^T\mathbf{h} + \overline{\theta} - \mathbf{x} \quad (42)$$

where $\underline{\mathbf{W}}$ and $\overline{\mathbf{W}}$ are the weight matrix of the input-hidden layer and the weight matrix of the hidden-output layer. $\underline{\theta}$ and $\overline{\theta}$ are the bias of the input-hidden layer and the hidden-output layer. ε is the residual error vector. If the Gaussian distribution kernel is used, the following approximation of error entropy can be derived.

$$H(\varepsilon) \approx -\frac{1}{N_B}\sum_{\varepsilon_B \in B} \log\left\{\frac{1}{N_A}\sum_{\varepsilon_A \in A} G_\Psi(\varepsilon_B - \varepsilon_A)\right\} \quad (43)$$

$$G_\Psi(\varepsilon) = \frac{1}{\sqrt{(2\pi)^n \det(\Psi)}}\exp\left(-\tfrac{1}{2}\varepsilon^T\Psi^{-1}\varepsilon\right)$$

$$\varepsilon_A = \overline{\mathbf{W}}^T\mathrm{F}\left(\underline{\mathbf{W}}^T\mathbf{x}_A + \underline{\theta}\right) + \overline{\theta} - \mathbf{x}_A \quad (44)$$

$$\varepsilon_B = \overline{\mathbf{W}}^T\mathrm{F}\left(\underline{\mathbf{W}}^T\mathbf{x}_B + \underline{\theta}\right) + \overline{\theta} - \mathbf{x}_B \quad (45)$$

where ε_A and ε_B are the samples belonging to sample dataset $\mathbf{A}$ and $\mathbf{B}$. Ψ is the covariance matrix of Gaussian distribution density function which can be estimated from samples (Bell and Sejnowski, 1995). By using the gradient descending algorithm similar to common back propagation algorithm, Equation (43) can be derived by a pseudo argument Λ as

$$\begin{aligned} \frac{\partial}{\partial\Lambda}H(\varepsilon) &\approx -\frac{1}{N_B}\sum_{\varepsilon_B\in B}\left\{\frac{\sum_{\varepsilon_A\in A}\frac{\partial G_\Psi(\varepsilon_B-\varepsilon_A)}{d\partial\Lambda}}{\sum_{\varepsilon_A\in A}G_\Psi(\varepsilon_B-\varepsilon_A)}\right\} \\ &= \frac{1}{N_B}\sum_{\varepsilon_B\in B}\sum_{\varepsilon_A\in A}\left\{\frac{G_\Psi(\varepsilon_B-\varepsilon_A)}{\sum_{\varepsilon_A\in A}G_\Psi(\varepsilon_B-\varepsilon_A)}\frac{\partial}{\partial\Lambda}\left\{\frac{1}{2}D_\Psi(\varepsilon_B-\varepsilon_A)\right\}\right\} \end{aligned} \quad (46)$$

where $D_\Psi(\varepsilon_B-\varepsilon_A) = (\varepsilon_B-\varepsilon_A)^T\Psi^{-1}(\varepsilon_B-\varepsilon_A)$. Since there is the fact that

$$\frac{\partial\left(\varepsilon^T\Psi^{-1}\varepsilon\right)}{\partial\varepsilon} = 2\Psi^{-1}\varepsilon \quad (47)$$

Therefore, if let Λ be the weight matrix $\underline{\mathbf{W}}$, the derivative of error entropy to $\underline{\mathbf{W}}$ can be expressed as

$$\begin{aligned} \frac{\partial}{\partial\underline{\mathbf{W}}}D_\Psi(\varepsilon_B-\varepsilon_A) &= \frac{\partial\left\{D_\Psi(\varepsilon_B-\varepsilon_A)\right\}}{\partial(\varepsilon_B-\varepsilon_A)}\frac{\partial(\varepsilon_B-\varepsilon_A)}{\partial\underline{\mathbf{W}}} \\ &= 2\Psi^{-1}(\varepsilon_B-\varepsilon_A)\overline{\mathbf{W}}^T\left\{\mathrm{F}'\left(\underline{\mathbf{W}}^T\mathbf{x}_B+\underline{\theta}\right)\mathbf{x}_B-\mathrm{F}'\left(\underline{\mathbf{W}}^T\mathbf{x}_A+\underline{\theta}\right)\mathbf{x}_A\right\} \end{aligned}$$

Similarly, the derivatives of Equation (43) to $\overline{\mathbf{W}}$, $\underline{\theta}$ and $\overline{\theta}$ can also be obtained.

6. CONCLUSIONS

PCA is the most extensively applied component extraction technique with the optimal performance in the sense of MSE. However, MSE which only considers the second-order statistic is a suitable criterion for measuring the uncertainty of a linear and Gaussian stochastic process. But for the nonlinear and non-Gaussian system, entropy is an alternative criterion for measuring the uncertainty. In this paper MSE and minimum error entropy (MEE) based component extraction techniques are described in detail, and the equivalence between MSE and MEE criteria is also analyzed from information theory.

Acknowledgements: This work is supported by NSFC of China (60474050 and 60534010) and the 111 project (B08015) with Northeastern University of China where the second author is affiliated to. These are gratefully acknowledged.

REFERENCES

Achlioptas, D. and McSherry, F. (2001), Fast computation of low rank matrix approximations, *Proceedings of the thirty-third annual ACM symposium on Theory of computing*. (New York: ACM Press), .611-618.

Bell, T. and Sejnowski, J. (1995): An information-maximization approach to blind separation and blind deconvolution, *Neural Computation* **27**, 1129-1159.

Diamantaras, K.I. and Kung, S.Y. (1996): *Principal Component Neural Networks: Theory and Applications*. New York : John Wiley & Sons,.

Girolami, M. **(**1999): *Self-Organising Neural Networks : Independent Component Analysis and Blind Source Separation*, London: Springer -Verlag,.

Guo, Z. H. and Wang, H**.** (2004): A modified PCA based on the minimum error entropy, *Proc. ACC'04*, Boston, USA,.

He, Q. Peter, Wang, J. and Qin, S. J. **(**2004): A new fault diagnosis method using fault directions in Fisher discriminant analysis, *Texas-Wisconsin Modeling and Control Consortium Tech. Report*

Hyvärinen, A., Karhunen, J. and Oja, E. **(**2001): *Independent Component Analysis*, New York: John Wiley & Sons.

Jackson, J. E. and Mudholkar, G. S. (1979): Control procedures for residuals associated with principal component analysis, *Technometrics* **21**, 341-349.

Johansson, K. H. (2000): The quadruple-tank process: a multivariable laboratory process with an adjustable zero. *IEEE Trns. Control Systems Technology*, **8**, 456-465.

Jolliffe, I.T. **(**2002): *Principal Component Analysis*, 2nd edition, New York: Springer-Verlag,.

Linsker, R. (1988): Self-Organization in a Perceptual Network, *IEEE Computer*, **21**, 105-117.

MacGregor, J. F. and Koutodi, M. (1995): Statistical process control of multivariate processes. *Control Engineering Practice* **3**, 403-414.

Nomikos, P. and MacGregor, J. F. (1995): Multivariate SPC charts for monitoring batch process. *Technometrics*, **37**, 403-414.

Parzen, E. **(**1962): On the estimation of a probability density function and the mode, *Annals of Mathematical Statistics*, **33l**, 1065-1076.

Taylor J. and Plumbley M. **(**1993), *Information Theory and Neural Networks in Mathematical Approaches to Neural Networks*, J. Taylor, (Ed.), Elsevier Science Publication., Amsterdam, 307-340.

Valle, S., Li, W. and Qin, S. J. (1999): Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Industrial and Engineering Chemistry Research*, **38**, 4389-4401.

Wang H. (2002): Minimum entropy control of non-Gaussian dynamic stochastic systems. *IEEE Trns. Automatic Control*, **47**, 398-403.