Published in final edited form as:
Conf Proc IEEE Eng Med Biol Soc. 2013 ; 2013: 3969–3972. doi:10.1109/EMBC.2013.6610414.

Emphysema Classification Based on Embedded Probabilistic PCA

Teresa Zulueta-Coarasa[1], **Sila Kurugol**[3], **James C. Ross**[3], **George G. Washko**[2], and **Raúl San José Estépar**[3]

Raúl San José Estépar: rjosest@bwh.harvard.edu

[1]Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Ontario

[2]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA

[3]Laboratory of Mathematics in Imaging, Brigham and Women's Hospital, Boston, MA

Abstract

In this article we investigate the suitability of a manifold learning technique to classify different types of emphysema based on embedded Probabilistic PCA (PPCA). Our approach finds the most discriminant linear space for each emphysema pattern against the remaining patterns where lung CT image patches can be embedded. In this embedded space, we train a PPCA model for each pattern. The main novelty of our technique is that it is possible to compute the class membership posterior probability for each emphysema pattern rather than a hard assignment as it is typically done by other approaches. We tested our algorithm with six emphysema patterns using a data set of 1337 CT training patches. Using a 10-fold cross validation experiment, an average recall rate of 69% is achieved when the posterior probability is greater than 75%. A quantitative comparison with a texture-based approach based on Local Binary Patterns and with an approach based on local intensity distributions shows that our method is competitive. The analysis of full lungs using our approach shows a good visual agreement with the underlying emphysema types and a smooth spatial relation.

I. Introduction

Chronic Obstructive Pulmonary Disease (COPD) is an irreversible lung condition that involves different diseases of the airways and parenchyma[1]. This group of diseases are expected to be one of the major causes of morbidity and the third cause of mortality by 2020. Emphysema is one of the main pathophysiological manifestations of COPD, which can be defined as the destruction of the pulmonary alveoli walls implying an enlargement of the air spaces in the lung parenchyma [2]. Morphologically, most authors distinguish between three types of emphysema, centrilobular emphysema that affects the respiratory bronchioles, panlobular emphysema that implies the destruction of the whole acinus and paraseptal emphysema that is morphologically similar to the other two types but occurs by definition near the pleura. In this article we consider six patterns of interest: normal tissue (NT), paraseptal emphysema (PS), panlobular emphysema (PL) and three subtypes of centrilobular emphysema: mild, moderate and severe (CL1/CL2/CL3).

Computed Tomography has been used by clinicians to assess emphysema as CT findings show a high correlation with the real extent of the disease [2]. However, visual scoring and interpretation of the images are subjective and time consuming so different approaches for automatic emphysema quantification have been proposed. The primary technique used for the detection and objective quantification of emphysema is based on lung density or densitometry [3] while more recently developed approaches are based upon textural analysis. These methods combine features extracted from co-occurrence matrices [4], local

binary patterns (LBP) [5] or multi-resolution features obtained from filter banks [6], [7]. A simpler alternative based on kernel density estimation (KDE) [8] has been proposed recently. A different approach to this effort may be based on manifold learning. The data of interest lies on an embedded nonlinear manifold within the higher-dimensional image space. For example, this approach has been used successfully to recognize faces [9].

In this article we propose a novel approach to classify different patterns of emphysema based on a probabilistic interpretation of the manifold in which each pattern is embedded. Our main goal is not to propose a hard classifier; emphysema assessment is a complex task that involves a large inter-subject variability. Rather, we propose a method that computes the class membership posterior probabilities for each emphysema patterns. This probabilistic framework may provide both clinicians and emphysema quantification approaches with additional information to handle the uncertainty associated to this problem. To achieve this, we implement Probabilistic Principal Component Analysis (PPCA) [10] preceded by generalized Linear Discriminant Analysis (LDA) [11], a step designed to find the most discriminative lower dimensional space in which to apply PPCA. Probabilistic manifold approaches have been also proposed for the problem of face recognition elsewhere [12], however our approach is unique in the use of a supervised embedding step based on LDA and a formal probabilistic extension of PCA. To evaluate the performance of our approach we use 10 fold cross-validation schemes in a data set of 1337 emphysema samples obtained from 267 COPD subjects. Also a comparison with LBP [5] and with KDE [8] is carried out.

II. Methods

In this section, we will present our method (see Fig. 1). First, we perform an initial dimensionality reduction using a global PCA. Next, we find an embedding space for each emphysema pattern versus the rest where the discriminative information under projection is maximal using a generalized LDA [11]. Finally, we compute a PPCA model in the embedded space to obtain a class membership probability for the input image sample. PPCA is a linear manifold learning technique derived from a density estimation perspective.

A. Global linear embedding: PCA-step

Before applying LDA, a dimensionality reduction step is performed by means of PCA. Let $\mathbf{X}^k=[\mathbf{x}_1^k, \mathbf{x}_2^k \cdots \mathbf{x}_{N_k}^k]$, $k \in [1 \ldots N_c]$ be a set of observed d-dimensional data vectors where N_c is the number of classes (patterns) and N_k is the number of samples per class. In our case, the feature vectors are 31×31 pixel image patches arranged as a column vector. Using all the training samples, PCA computes a projection matrix $\mathbf{Q}_{PCA}$ that reduces the dimension from d to d_{PCA} with $d_{PCA} < d$ such as $\mathbf{Z}^k = \mathbf{Q}_{PCA}\mathbf{X}^k$. The basis of $\mathbf{Q}_{PCA}$ can be interpreted as a set of linear operators.

B. Supervised embedding: LDA-step

The aim of this step is to find the embedded space that best discriminates the k pattern from the others. Fisher Linear Discriminant Analysis performs a dimensionality reduction preserving as much class discriminatory information as possible. This technique fits very well to our problem for two reasons. First, it is a supervised method, so we can take advantage of our data labeling. Second, the resulting projections have a compact, monomodal distribution that can be efficiently captured by PPCA (as can be seen in the example given in Fig. 2). Instead of using the traditional algorithm, which only finds one discriminant vector, we use the generalization proposed by Duchene and Leclercq that permits to find n discriminant vectors [11]. For each emphysema pattern, k, a LDA model is defined using two classes, $C_1 = \mathbf{Z}^k$ and $C_2 = \{\mathbf{Z}^1 \ldots \mathbf{Z}^{k-1}\mathbf{Z}^{k+1} \ldots \mathbf{Z}^{N_k}\}$ such that the ratio

$$J^k = \frac{\mathbf{q}^{k^t}{}_n \mathbf{S}_B^k \mathbf{q}_n^k}{\mathbf{q}^{k^t}{}_n \mathbf{S}_W^k \mathbf{q}_n^k} \quad (1)$$

is maximized, where $\mathbf{S}_B^k$ is the between-class covariance matrix, $\mathbf{S}_W^k$ is the within-class covariance matrix for the classes C_1 and C_2.

The method proposed in [11] maximizes J^k by computing a linear subspace $\mathbf{Q}_{LDA}^k = [\mathbf{q}^{k^t}{}_1, \mathbf{q}^{k^t}{}_2 \cdots \mathbf{q}^{k^t}{}_{d_{LDA}}]^t$ that can be used to project each image vector onto the most discriminant embedded space for pattern k of dimensionality d_{LDA}.

C. Probabilistic Emphysema Classification: PPCA-step

PPCA derives PCA in the maximum likelihood framework allowing to calculate posterior class-membership probabilities in a formal way. The approach is described in detail in [10].

Using a Gaussian latent variable model, it is possible to compute the likelihood of an input vector **y** for a given emphysema pattern *k*.

$$p(\mathbf{y}|k) = (2\pi)^{-d_{LDA}/2} \left|\mathbf{C}^k\right|^{-1/2} e^{-\frac{(\mathbf{y}-\mu^k)^t \mathbf{C}^{\mathbf{k}-1}(\mathbf{y}-\mu^k)}{2}} \quad (2)$$

where $\mu^k = \frac{1}{N}\sum_{n=1}^{N_k} \mathbf{Q}_{LDA}^k \mathbf{Q}_{PCA}(\mathbf{x}_n^k - \overline{x})$ is the mean of the data for pattern k and $\bar{x}$ is the global mean of the data.

$$\mathbf{C}^k = {\sigma_k}^2 \mathbf{I} + \mathbf{W}^k \mathbf{W}^{k^t} \quad (3)$$

is the model covariance for pattern k where ${\sigma_k}^2 = \frac{1}{d_{LDA} - d_{PPCA}} \sum_{j=d_{PPCA}+1}^{d_{LDA}} \lambda_j$ is the noise variance and λ_j are the smallest eigenvalues of the sample covariance for each class

$$\mathbf{S}^k = \frac{1}{N}\sum_{n=1}^{N_k} \mathbf{Q}_{LDA}^k \mathbf{Q}_{PCA}(\mathbf{x}_n^k - \mu^k)(\mathbf{x}_n^k - \mu^k)^t. \quad (4)$$

$\mathbf{W}^k = \mathbf{U}_{d_{PPCA}}^k (\mathbf{\Lambda}_{d_{PPCA}}^k - {\sigma_k}^2 \mathbf{I})^{1/2} \mathbf{R}$ is the weight matrix for each pattern, where the d_{PPCA} column vectors of the $d_{LDA} \times d_{PPCA}$ matrix $\mathbf{U}_{d_{PPCA}}^k$ are eigenvectors of $\mathbf{S}^k$, with corresponding eigenvalues in the $d_{PPCA} \times d_{PPCA}$ diagonal matrix $\mathbf{\Lambda}_{d_{PPCA}}^k$ and **R** is an arbitrary $d_{PPCA} \times d_{PPCA}$ orthogonal rotation matrix.

The posterior probability for the emphysema pattern *k* given the data **y** is defined by means of the Bayes' rule as

$$p(k|\mathbf{y}) = \frac{p(\mathbf{y}|k)\Pi_k}{\sum_{j=1}^{N_c} p(\mathbf{y}|j)\Pi_j} \quad (5)$$

where $\Pi_i = \frac{N_k}{N_T}$ are the priors and N_T is the total number of samples in the training set.

Figure 2 shows an example of how the method works when the data is embedded in a two-dimensional space using our training set. 20% of the samples for each class were segregated and projected with the models computed from the rest of the available samples. LDA is successful at defining an embedded subspace that separates a given pattern k from the rest. Then, PPCA estimates a likelihood density function (represented with isolines) that can be used to evaluate new data (represented as squares). The separability achieved by LDA is dependent on the initial dimensionality reduction performed by PCA (higher d_{PCA} implies better separability), however the generalization of the LDA model to new examples is affected by high d_{PCA} values.

III. Results

Emphysema Database

In our experiments we utilized 1337 training samples that were labelled by an expert. The distribution of samples per pattern was: NT=370, PS=184, PL=148, CL1=287, CL2=178, CL3=178. The samples were selected from a group of 267 subjects scanned across 16 different institutions as part of the COPDGene study. On average, the expert labeled six samples per patient at random based on prototypic expression of disease and without any prior spatial correlation. As such we can consider that the samples are independent representations of disease regardless of the patient that were selected from. The expert performed a second review to evaluate the consistency of the assignments and samples that were non-consistently labeled were discarded. The spatial size of the samples was chosen to fit the physical extent of emphysema within the secondary lobule corresponding to 31 × 31 pixels patches (d = 961). Prior to the application of our method, each image pixel was normalized by means of the z-score using the global mean and standard deviation of all the training pixel values.

Parameter selection

Our approach involves three dimensionality reduction steps, PCA, LDA and PPCA, and the embedding dimensionality in each stage of the method is a free parameter. We set d_{PCA}, d_{LDA} and d_{PPCA} to optimize the classification accuracy defined as the distance to the perfect classifier on the training set using a nested 10-fold cross validation [13] with a grid search ($d_{PCA} \in [20, 100]$, $d_{LDA} \in [10, 20]$ and $d_{PPCA} \in [2, 10]$). In a nested cross-validation, the training set for each fold is used in a new cross validation experiment to set the optimal parameters for that fold. This reduces the bias due to parameter optimization. When using the full training set, the optimal parameters were: $d_{PCA} = 22$, $d_{LDA} = 17$ and $d_{PPCA} = 8$. These values are included as a reference for the optimality range.

Classification performance

To assess the classification performance of our approach, we used a nested 10-fold cross validation as described before and we carried out a comparison with the LBP method [5] and with KDE [8] using the optimal parameters described in their paper. Both methods employ a kNN classifier that does not provide posterior probability estimates for the classifier assignments. In our method, we used a Maximum a Posterior (MAP) criteria to compute the confusion matrix. The assignments were done at different confidence levels (25%,50%,75% and 90%) given by the upper threshold on the MAP probability, i.e. $\text{EPPCA}_{75\%}$ uses those test samples that were assigned a posterior probability greater than 0.75. Precision, sensitivity and specificity are shown in Table I. We can see how our method is competitive with LBP and KDE at a confidence level of 25%, i.e when all the samples are used to compute the confusion matrix. Our method improves its accuracy when a higher confidence level is used, reflecting the important information that the posterior probability conveys.

Full lung emphysema classification

For a more exhaustive evaluation of our approach we computed a full lung classification in subjects with different patterns of disease. For a given CT slice, we applied our approach to each voxel and computed the class membership probabilities for each emphysema type. Fig. 3 shows the probability maps for each emphysema pattern for three subjects. It is worth noting how paraseptal is mostly present in late stages diseases while mild centrilobular is present for the mild stages. Paraseptal is confined to the pleural region as it should be expected by the histopathology of this disease. The normal smoker lung shows moderate level (mid probabilities) of mild centrilobular disease (CL1) that might suggest the impact of smoking in the lung parenchyma even when emphysema has not been fully manifested.

IV. Discussion and Conclusions

In this paper we present a new approach to quantify different emphysema patterns, based on a optimal embedded PPCA. Our approach to the emphysema classification problem is novel in that we capture the underlying data manifold for each pattern (or class) in a probabilistic fashion using a supervised embedding technique based on LDA. The embedded space in which the probabilistic data model is learnt is computed by means of a generalized LDA that is trained for each tissue type to maximize the inter-class covariance between that emphysema type and the rest of types. Our method shows a performance that is comparable to current techniques.

The need for a initial dimensionality reduction step based on PCA is twofold. First, when the dimensionality of the data is bigger than the number of samples, the within-class covariance matrix J^k is likely to be singular. Also, we have noticed that this step is necessary to achieve a balance between the discriminative power of LDA and the generalization of the model when projecting new samples. Although in this paper we have used a linear dimensionality reduction approach, more general non-linear dimensionality reduction methods could also be explored and applied.

The posterior probability information provided by this method can be used in multiple ways. Certainties about the most likely labels can be provided to guide the emphysema quantification stage. Additionally, advanced method based on Markov chain models can be used to relax the probability assignments taking into account priors about disease progression between stages and spatial relationships of disease.

Acknowledgments

This work has been funded by grants from the National Institutes of Health 1R01HL116931-01. RSJE is supported by K25 HL104085 and GW by K23 HL089353. The COPDGene Study is funded by 2R01HL089897-06A1 and 2R01HL089856-06A1. The authors would like to thank to all the COPDGene Investigators for their contributions.

References

1. MacNee W, Tuder RM. New paradigms in the pathogenesis of chronic obstructive pulmonary disease I. Proceedings of the ATS. Sep; 2009 6(6):527–531.
2. Thurlbeck WM, Müller NL. Emphysema: definition, imaging, and quantification. AJR. 1994; 163(5):1017–1025. [PubMed: 7976869]
3. Müller NL, et al. Density mask. An objective method to quantitate emphysema using computed tomography. Chest. 1988; 94(4):782–787. [PubMed: 3168574]
4. Haralick RM, et al. Textural Features for Image Classification. IEEE Trans Systems Man and Cybernetics. 1973; 3
5. Sorensen L, et al. Quantitative Analysis of Pulmonary Emphysema Using Local Binary Patterns. TMI. 2010; 29:559–569.

6. Sluimer IC, et al. Computer-aided diagnosis in high resolution CT of the lungs. Medical physics. 2003; 30
7. Depeursinge A, et al. Multiscale Lung Texture Signature Learning Using the Riesz Transform. MICCAI. 2012:7512.
8. Mendoza, CS., et al. Emphysema quantification in a multi-scanner HRCT cohort using local intensity distributions. ISBI 2012; Barcelona. May 2012; p. 474-477.
9. Belhumeur P, et al. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. PAMI. 1997; 19:711–720.
10. Tipping ME, Bishop CM. Mixtures of probabilistic principal component analyzers. Neural computation. 1999; 11(2):443– 482. [PubMed: 9950739]
11. Duchene J, Leclercq S. An optimal transformation for discriminant and principal component analysis. PAMI. 1988; 10(6):978–983.
12. Moghaddam B. Principal Manifolds and Probabilistic Subspaces for Visual Recognition. PAMI. 2002; 24(6)
13. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC bioinformatics. 2006; 7:91. [PubMed: 16504092]

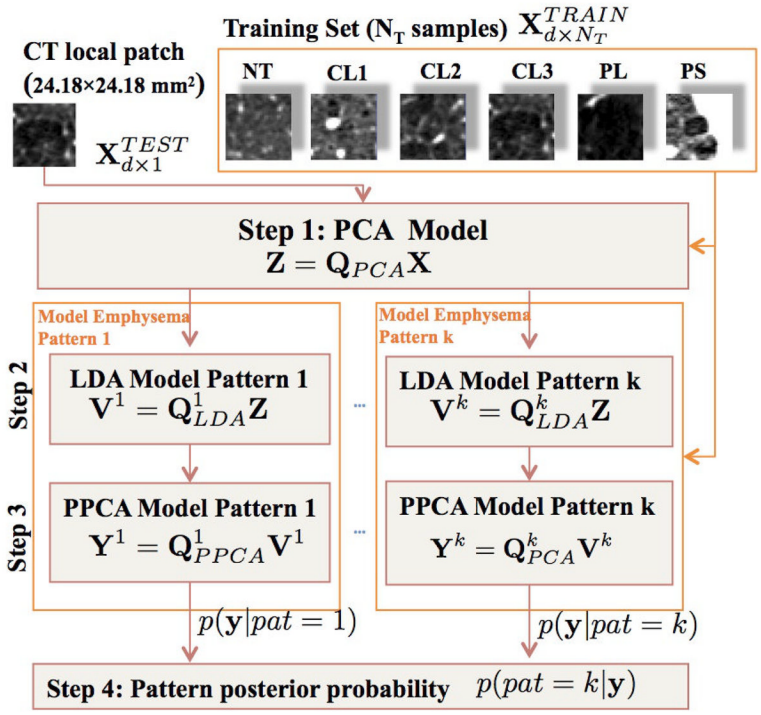

Fig. 1.

Schematic view of our method. For each emphysema pattern a LDA and a PPCA model capture is trained to capture the likelihood of each pattern.

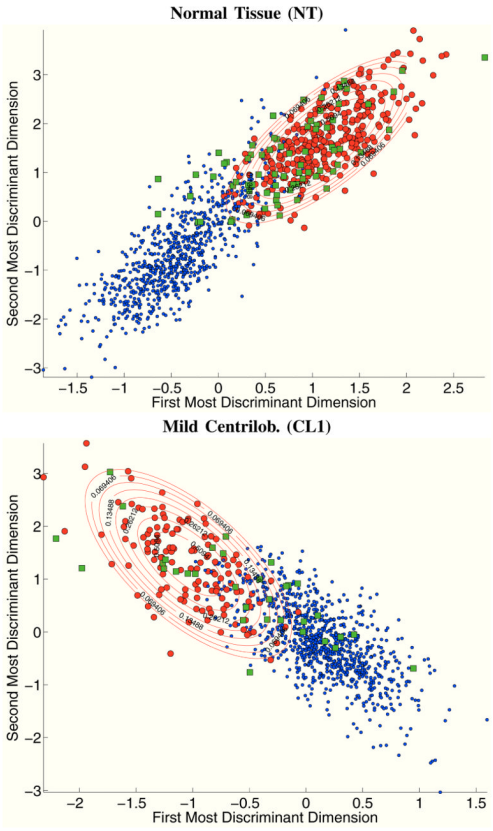

Fig. 2.

An example of the LDA feature space for $d_{PCA} = 400$ and $d_{LDA} = 2$ for two tissue types. For both normal tissue (top) and mild centrilobular (bottom) emphysema, the manifold learnt by LDA properly separates those classes from the other class samples using a training set corresponding to the 80% of the available samples. The isolines represent the likelihood for the data given by the PPCA model. Green squares represent the testing data for each class projected in the model trained for that class (red circles). Blue circles represent the rest of the data points projected in the LDA-derived embedding space for that model. It is worth noting how the LDA model is able to project new data into the proper embedded space that is in turn captured by PPCA.

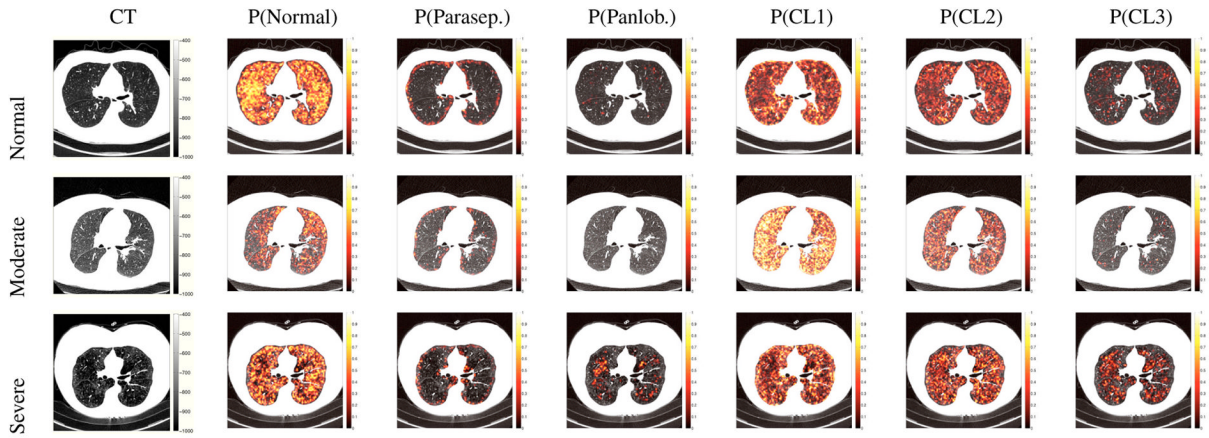

Fig. 3.

Class membership Posterior Probability maps for three subjects with different disease severity: normal smoker (top), moderate disease (middle) and severe disease (bottom). As disease progresses, the posterior probability increases in the moderate and severe centrilobular classes (CL1 and CL2) and the panlobular class. It is important to note that the paraseptal posterior probability only shows signal in the pleural interface as it should be expected.

TABLE I

Classification performance metrics derived from the confusion matrix. Our method (EPPCA) classifies according to a MAP criteria using different confidence levels and is compared with LBP [5] and KDE [8].

Method	**Precision**						**Sensitivity (Recall)**						**Specificity**						**Total mean(± std)**		
	NL	**PS**	**PL**	**CL1**	**CL2**	**CL3**	**NL**	**PS**	**PL**	**CL1**	**CL2**	**CL3**	**NL**	**PS**	**PL**	**CL1**	**CL2**	**CL3**	**Prec.**	**Sens.**	**Spec.**
LBP	0.90	0.82	0.75	0.32	0.63	0.48	0.80	0.96	0.74	0.43	0.59	0.47	0.94	0.96	0.96	0.88	0.87	0.90	0.65(±0.22)	0.66(±0.20)	0.92(±0.04)
KDE	0.89	0.85	0.78	0.34	0.64	0.45	0.81	0.97	0.71	0.42	0.58	0.52	0.94	0.96	0.96	0.89	0.88	0.90	0.66(±0.22)	0.67(±0.20)	0.92(±0.04)
$EPPCA_{25\%}$	0.76	0.95	0.82	0.52	0.46	0.51	0.84	0.76	0.70	0.51	0.59	0.45	0.87	0.99	0.97	0.91	0.83	0.90	0.67(±0.20)	0.64(±0.15)	0.91(±0.06)
$EPPCA_{50\%}$	0.77	0.95	0.83	0.52	0.48	0.50	0.85	0.76	0.71	0.51	0.60	0.47	0.87	0.99	0.97	0.91	0.85	0.90	0.68(±0.20)	0.65(±0.15)	0.92(±0.05)
$EPPCA_{75\%}$	0.83	0.98	0.88	0.52	0.50	0.47	0.90	0.79	0.74	0.64	0.58	0.47	0.91	0.99	0.98	0.93	0.88	0.91	0.70(±0.22)	0.69(±0.15)	0.93(±0.04)
$EPPCA_{90\%}$	0.80	0.99	0.92	0.60	0.56	0.43	0.93	0.81	0.77	0.73	0.61	0.54	0.92	0.99	0.98	0.95	0.92	0.92	**0.72**(±0.22)	**0.73**(±0.14)	**0.95**(±0.03)