

Nonnegative Principal Component Analysis for Proteomic Tumor Profiles

Xiaoxu Han*

Abstract

Identifying cancer molecular patterns with high accuracy from high-dimensional proteomic profiles presents a challenge for statistical learning and oncology research. In this study, we develop a nonnegative principal component analysis and propose a nonnegative principal component analysis based support vector machine with a sparse coding to conduct effective feature selection and high-performance proteomic pattern classification. We demonstrate the superiority of our algorithm by comparing it with six peer algorithms on four benchmark proteomic tumor profiles under 100 trials of 50% holdout cross validations. We also rigorously show that the over-fitting problem associated with support vector machines can be overcome by nonnegative principal component analysis with exceptional sensitivities and specificities. Moreover, we illustrate that nonnegative principal component analysis can be employed to capture meaningful biomarkers.

1 Introduction.

As a promising early cancer diagnosis technique in oncology, mass spectrometry (MS) generates protein expression data from a patient's serum, plasma or urine samples. These expressions data are the signatures of cancers at a molecular level that reflect protein activity patterns in different types of cancerous or precancerous cells. Although there is an urgent need to predict cancer molecular patterns with high accuracy to support clinical decisions, it is still a challenge for oncologists and computational biologists to achieve a high-performance classification due to the special characteristics of mass spectral data.

A mass spectral profile is generally characterized by large or even huge dimensionalities [1]. It can be represented by a $d \times n$ matrix, each column of which represents intensity values of all biological samples in investigation at a mass charge ratio (m/z); each row of which represents ion-intensity values of a single biological sample at different m/z values. The protein

expression at each m/z ratio can be called a pseudo-gene because it is similar to a gene in a gene expression profile. Generally, the total number of m/z ratios is in the order of $10^4 \sim 10^6$, and the total number of biological samples is on the magnitude of hundreds, i.e., the number of variables is much greater than the number of biological samples. Although there are a large number of m/z ratios in a mass spectral profile, only a small number of them have meaningful contributions to the data variations. Moreover, the high-dimensional data are not noise-free because the raw data contains systematic noise and preprocessing algorithms may not remove it completely.

Many feature selection algorithms are employed to reduce data dimensions, remove noise, and extract meaningful features before further classification. These algorithms include basic two-sample t-tests [2], principal component analysis (PCA)[3], independent component analysis (ICA) [4], nonnegative matrix factorization (NMF) [5], and their different variants [6, 7]. Principal component analysis (PCA) [8] may be the most employed among them for its simplicity. It projects data in an orthogonal subspace generated by the eigenvectors of the data covariance matrix. The maximum variance direction-based subspace spanning guarantees the least information loss in the feature selection.

However, as a holistic feature-selection algorithm, PCA can only capture global features instead of local features [5-6]. The global and local features contribute to the global and local characteristics of data, which are responsible to interpret the global and local behavior of data respectively. The local features are more difficult to be extracted than the global features for their frequencies. However, local feature selection may be crucial in attaining high-accuracy cancer diagnosis. For example, some benign tumor samples pathologically may display very similar global characteristics with malignant tumor samples but with different local characteristics. Obviously, selecting local features will be a key in distinguishing the sample sharing similar global features.

The standard PCA by nature is a global feature selection algorithm. Each principal component in PCA

*Department of Mathematics and Bioinformatics, Eastern Michigan University, Ypsilanti MI 48197 USA. Email: xhan1@emich.edu

contains some levels of global characteristics of data and receives contributions from all input variables in the linear combinations. Changes in one variable will inevitably affect all loading vectors globally. In addition to causing difficulties in interpreting each principal component intuitively, its global feature-selection mechanism prevents a high-accuracy prediction in the following classification, because only the features interpreting global characteristics are involved in training a learning machine such as a support vector machine (SVM) or its extensions [9, 10, 11, 12]. The redundant global features in training will decrease the generalization of the learning machine and increase the risk of misclassifications or over-fitting.

There are two major reasons for PCA's global feature selection mechanism. One is that data representation in PCA is not purely-additive, where the linear combination to calculate each PC includes both positive and negative weights. The positive and negative weights are likely to partially cancel each other in the linear combinations. The weights representing contributions from local features are more likely to be cancelled because of their frequencies. The partial-cancellation directly causes missing captures of local features in each loading vector. Another is that PCA lacks some level sparse representation where each loading vector receives contributions from all input variables in the linear combination. Any changes in one variable will globally affect all loading vectors.

We overcome the global feature selection mechanism by adding nonnegative and sparse constraints on PCA. Imposing nonnegativity constraints on PCA that restricts all entries of the input data and loading vector nonnegative can remove the partial-cancellations and make data representation consisting of only additive components. Moreover, adding sparse constraints on PCA (i.e., increasing number of zeros for each loading vector) can highlight important variables' contributions that are more sensitive than the others in distinguishing cancer and normal samples for the sake of biomarker discovery. Mathematically, it enhances data locality in the feature selection by decomposing the support sets of the input data into as a set of small locally-compact ones.

In this study, we present a nonnegative principal component analysis (NPCA) algorithm and propose a nonnegative principal component analysis based support vector machine (NPCA-SVM) for efficient proteomic pattern classification. We demonstrate our algorithm's superiority by comparing it with other peer classification algorithms on four benchmark profiles. This paper is organized as follows. Section 2 presents the nonnegative principal component analysis and nonneg-

ative principal component analysis based support vector machine. Section 3 compares the NPCA-SVM algorithm with six peer algorithms. Section 4 gives a rigorous analysis on the special characteristics of the SVM over-fitting and points out that it can be overcome by nonnegative principal component analysis. Section 5 presents a NPCA-based biomarker discovery algorithm for two mass spectral datasets. Finally, we discuss the limitations and potential improvements on nonnegative principal component analysis in proteomic pattern predictions.

2 Nonnegative Principal Component Analysis.

Nonnegative PCA can be viewed as an extension of the classic PCA algorithm by imposing it with nonnegativity constraints to capture data locality in the feature selection. Let $X = [x_1, x_2 \dots x_n]$, $x_i \in \mathbb{R}^d$, be a zero mean dataset, the nonnegative PCA can be formulated as a constrained optimization problem to find maximum variance directions under nonnegative constraints as follows.

$$(2.1) \quad \max J(U) = \frac{1}{2} \|U^t X\|_F^2, \text{ st. } U^t U = I, U \geq 0$$

where $U = [u_1, u_2, \dots u_k]$, $k \leq d$, is a set of nonnegative PCs. The square Frobenius norm for a matrix A is defined as $\|A\|_F^2 = \sum_{i,j} a_{ij}^2 = \text{trace}(AA^t)$. The orthonormal constraint under nonnegativity in equation (2.1) requires each loading vector contain only one nonnegative entry. Although it may be theoretically attractive, it would be too harsh in a practical proteomic pattern analysis. It is possible that several or more key pseudo-genes are essential biomarkers in cancer detection and their contributions to each PC should not be limited to only one entry. Thus, the quadratic programming problem can be further relaxed by using a parameter $\alpha \geq 0$ to control the orthonormal degree of each loading vector as follows,

$$(2.2) \quad \max J(U, \alpha) = \frac{1}{2} \|U^t X\|_F^2 - \alpha \|I - U^t U\|_F^2$$

The principal component matrix U is a near-orthonormal nonnegative matrix: $U^t U \approx I$. Calculating the gradient of the objective function with respect to U , we have the following gradient learning scheme for the relaxed quadratic programming problem,

$$(2.3) \quad U(t+1) = U(t) - \eta(t) \frac{\nabla_U J(t)}{\|\nabla_U J(t)\|}, U \geq 0$$

where $\eta(t)$ represents the t time level iteration step size and $\nabla_U J(t) = (U^t X)X^t + 4\alpha(I - U^t U)U^t$. We select $\eta(t) = 1$ to avoid an expensive trust region search. This is equivalent to finding the local maximum of a scalar

function $f(u_{sl})$ under the constraints $u_{sl} \geq 0$, $s = 1, 2 \dots d$, $l = 1, 2 \dots n$,

$$(2.4) \quad \max_{u_{sl} \geq 0} f(u_{sl}) = -\alpha u_{sl}^4 + c_2 u_{sl}^2 + c_1 u_{sl} + c_0$$

where c_2 , c_1 and c_0 are the parameters to be determined in the local optimum finding of $f(u_{sl})$, which is a set of cubic polynomial nonnegative root finding problems. The final principal component matrix U is a collection of nonnegative roots of equation (2.4). Calculating the stationary points and collecting the coefficients of u_{sl} and u_{sl}^2 , we have the following parameters,

$$(2.5) \quad c_2 = \frac{1}{2} \sum_{i=1}^n x_{si}^2 - \alpha \sum_{j \neq l} u_{sj}^2 - 2\alpha \sum_{t \neq s} u_{tl}^2 + 2\alpha$$

$$(2.6) \quad c_1 = \sum_{i=1}^n \sum_{t \neq s} x_{si} u_{tl} x_{ti} - 2\alpha \sum_{j \neq l} \sum_{t \neq s} u_{sj} u_{tl} u_{tj}$$

The coefficient $c_0 = -k\alpha$ does not have any contributions to the entries of the principal component matrix. Only coefficients c_2 and c_1 are involved in the nonnegative root finding. The nonnegative principal component analysis complexity is $O(dkn \times N)$, where N is the total iterations needed to meet the algorithm termination threshold $\|\nabla_U J(t)\| \leq 10^{-4}$ in the implementation. It is worthy to point that Zass and Shashua also proposed a similar approach to solve a nonlinear optimization problem induced by a nonnegative sparse PCA [13], where two penalty parameters are employed to control the orthonormality and sparseness of the PC matrix. Since there are still no robust algorithms available to determine the penalty parameters, an additional sparseness control parameter will increase the risk of algorithmic convergence difficulty with the increasing of the parameter values [14, 15]. It would be more desirable to achieve sparseness for each loading vector without increasing the difficulty of the algorithmic convergence.

2.1 Nonnegative Principal Component analysis based Support Vector Machine Algorithms. As a novel feature selection methodology to overcome the global nature of PCA and improve data locality, it is desirable to apply it in proteomic pattern classifications to attain high accuracies. In this section, we propose a nonnegative principal component analysis based support vector machine (NPCA-SVM) for mass spectral pattern prediction. The goal of the NPCA-based classification algorithm is to employ nonnegative principal component analysis to obtain a nonnegative representation of each sample in a low-dimensional, purely additive subspace spanned by meta-variables. A meta-variable is

a linear combination of the intensity values of the measured data points (pseudo-genes) for in a mass spectral profile. The nonnegative representation for each sample is denoted as a meta-sample that is the prototype of the original sample with low dimensionalities. Then, a classification algorithm, which is chosen as a support vector machine (SVM) [11] in this study, is applied to the meta-samples to gain classification information.

Theoretically, our algorithm is rooted in a NPCA-induced nonnegative matrix factorization (NMF) that we propose in this study. We outline the principle of this special nonnegative matrix factorization as follows. Let $X \in \mathbb{R}^{d \times n}$, $d \ll n$ be a nonnegative matrix, which is a mass spectral profile with d samples across n m/z ratios, and $U \in \mathbb{R}^{d \times d}$ be the nonnegative PC matrix before any further dimension selection, we project X^t into a purely-additive subspace generated by U , and obtain $X^t U = P$. Since the PC matrix U is a near-orthogonal matrix, it can be viewed as an orthogonal matrix to decompose the data matrix approximately: $X^t \sim P U^t$. The nonnegative matrices P and U^t equivalent to the basis matrix W and the feature matrix H in the classic NMF algorithm respectively, i.e., $X \sim WH$. Similarly, the decomposition rank r in the nonnegative matrix factorization is just the selected dimensionality in the nonnegative principal component analysis.

The nonnegative principal component analysis induced NMF can be further explained as follows. Each row of U is the corresponding meta-sample of each biological sample of X in a meta-variable space: $X_i^t \sim P U_i^t$. The meta-variable space is a subspace generated by columns of the basis matrix P , where each basis is a meta-variable. The meta-variable space is a purely-additive space and each variable can be represented as the nonnegative linear combination of meta-variables as $X_i^t = \sum_{j=1}^r U_{ij}^t P_j$, $1 \leq r \leq d$.

According to the observation that mass spectral data are naturally positive or can be converted to its corresponding nonnegative data easily, we have the following nonnegative principal component analysis based support vector machine (NPCA-SVM) for mass spectral data classification, Given an input dataset $X \in \mathbb{R}^{d \times n}$, nonnegative principal component analysis is used to obtain the low-dimensional but locality-preserved meta-samples $U \in \mathbb{R}^{d \times k}$, $k \leq d \ll n$ for all biological samples. To improve algorithmic generality and robustness, the normalized meta-samples $U = U/\|U\|_2$ work as input data for the support vector machine.

The algorithm can be officially described as follows. Given a protein expression training dataset consisting of d biological samples across n pseudo-genes and their label information: $\{x_i, c_i\}_{i=1}^d$ where $X = [x_1, x_2 \dots x_d]^t$, $x_i \in \mathbb{R}^n$, and $c = [c_1, c_2 \dots c_d]^t$, $c_i \in \{-1, 1\}$,

the NPCA-SVM algorithm finds the meta-samples $U = [u_1, u_2 \dots u_d]^t$, $U \in \mathbb{R}^{d \times k}$, $k \leq d \ll n$, by the steepest descent method given in equation (2.3). Then, an optimal separating hyperplane $O_h : w^T u + b = 0$ in \mathbb{R}^d is computed to attain the maximum margin between the $'-1'$ and $'+1'$ types of the meta-samples. This is equivalent to solving the following quadratic programming problem in \mathbb{R}^d ,

$$(2.7) \quad \begin{aligned} \min & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^d \xi_i \\ \text{s.t.} & c(w^T u_i + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, 2 \dots d \end{aligned}$$

Given an unknown type sample $x' \in \mathbb{R}^n$, the NPCA-SVM learning machine employs function $f(x') = \text{sign}(\sum_{i=1}^d \alpha_i c_i k(u_i, u') + b)$ to determine its class type, where $u, u' \in \mathbb{R}^d$ are the meta-samples of samples x and x' computed from nonnegative principal component analysis respectively. The vector $\alpha = [\alpha_1, \alpha_2 \dots \alpha_d] \geq 0$ is the solution the dual problem of equation (2.7), and $k(u_i, u')$ is a kernel function for the support vector machine that maps these meta-samples into a same-dimensional or high-dimensional feature space. We only focus on the linear and standard Gaussian ($'rbf'$) kernels due to their popularity [12].

We implement the nonnegative principal component analysis based support vector machine under 100 trials of 50% holdout cross validations (HOCV), where 100 sets of training and test datasets are generated randomly for each dataset. The algorithm performance is evaluated as the average classification performance of 100 trials. To improve computing efficiency, the PC matrix U in the nonnegative principal component analysis is cached from the previous trial and used as the initial point to compute the next principal component matrix in the computation.

2.2 Sparse-coding for nonnegative principal components

It would be ideal to improve the sparseness of the nonnegative principal component matrix without increasing the algorithmic convergence difficulties. Moreover, increasing the meta-samples sparseness contributes to enhancing algorithmic generalizations. We employ a sparse-coding approach to improve the sparseness for a meta-sample that is a nonnegative loading vector. The sparseness of a nonnegative vector $v = [v_1, v_2 \dots v_n]^t$, $v_i \geq 0$, $i = 1, 2 \dots n$ is defined as a ratio as follows:

$$(2.8) \quad \delta_v = \frac{\sqrt{n} - \|v\|_1 / \|v\|_2}{\sqrt{n} - 1}$$

A large sparseness value indicates less number of positive entries in the vector v . $\delta_v = 1$ or $\delta_v = 0$ indicate that there is only one entry or all entries are equal in v respectively. The sparse coding of a meta-sample

Table 1: Four serum mass spectral profiles

Dataset	Technology	#m/z	#Samples
Ovarian	SELDI-TOF Low	15142	91 'H'+162 'C'
Ovarian-qaqc	SELDI-TOF High	15000	95 'H' + 121 'C'
Liver	SELDI-QqTOF High	6107	176 'H' + 181 'C'
Colorectal	MADLI-TOF High	16331	48 'H' + 64 'C'

$u_i^t \in \mathbb{R}^{1 \times k}$, $i = 1, 2 \dots d$, $k \leq d \ll n$ seeks to find a corresponding nonnegative vector $v \in \mathbb{R}^{1 \times k}$ such that $\|v\|_1 = \|u_i^t\|_1$, $\|v\|_2 = \|u_i^t\|_2$, and δ_v equal to a specified sparseness value. In other words, for each loading vector u_i^t in the nonnegative PC matrix, the nearest nonnegative vector v on behalf of L_1 and L_2 distances is found to achieve a specified sparseness. Alternatively, it is equivalent to calculating the nonnegative intersection point between a hyperplane $\pi_1 : \sum_{j=1}^k v_j = \sum_{j=1}^k u_{ij}$ and a hypersphere $\pi_2 : \sum_{j=1}^k v_j^2 = \sum_{j=1}^k u_{ij}^2$ such that the sparseness is achieved as $\delta_v = (\sqrt{k} - \|u_i^t\|) / (\sqrt{k} - 1)$. This optimization problem can be solved in real-time by a traditional approach [15] or by a simple but efficient method given in [6].

It is interesting to discuss how to set the optimal parameter values for the orthonormal control α and the sparseness δ_v for each each nonnegative loading vector. Theoretically, the optimal value for α can be determined by constructing an increasing sequence $\{\alpha_k\}$ and solving corresponding quadratic programming problems in equation (2.2) [14]. The optimal parameter is a value $\alpha^* \in \{\alpha_k\}$ that achieves a maximum value for the equation (2.2). Obviously, this approach is computationally expensive. In practice, we select $\alpha \propto d$ because of $\|U^t U\| = I$ in the extreme case (U is an identity matrix), if there is no further sparse coding applied to loading vectors. If there is a sparse coding processing applied to each loading vector, we select $\alpha \propto \sqrt{d}$. Since data sparseness is a byproduct of the nonnegativity constraints in the equation (2.2), we usually select the sparseness degree for each nonnegative principal component $\delta_v \leq 0.5$.

3 Results.

Four serum proteomic datasets: ovarian, ovarian-qaqc (quality assurance/quality control), liver and colorectal, are included in this study [16-18]. They include one low resolution and three high resolution datasets, which are generated from three different profiling technologies: SELDI-TOF, SELDI-QaTOF and MADLI-TOF respectively. Table 1 provides the detailed information about the datasets, where 'Low'/'High' indicates 'high/low resolution' and 'H' and 'C' represent 'control' and 'cancer' samples respectively.

We conducted the following pre-processing steps [19-20] for each dataset: spectrum calibration, base-

line correction, smoothing, peak identification, intensity normalization, and peak alignments. In addition, we applied the standard two-sample t-test to conduct a basic feature selection to choose 3780, 2500, 3000 and 1000 most significant pseudo-genes for the ovarian, ovarian-qaqc, liver and colorectal data respectively before following classifications. The goal of this basic feature selection is to select approximately $10 \times n$ most significant features for each input dataset $X \in \mathbb{R}^{d \times n}$ before classification.

We compare our algorithm with its two siblings: support vector machines (SVM) and principal component analysis based support vector machine (PCA-SVM) under 100 trials of 50% HOCV. The PCA-SVM algorithm conducts SVM classification by projecting the testing data into the subspace spanned by the training data. In the NPCA-SVM algorithm, we set the relevant parameters $\alpha = 10$, $\delta_v = 0.20$, and $k = d - 1$ respectively. Table 2 shows the average performance of the three algorithms in terms of average classification rates $E(r_c)$, sensitivities $E(r_s)$, specificities $E(r_p)$, and their corresponding standard deviations on the four profiles.

We have the following observations from these classification results. 1). The SVM and PCA-SVM algorithms both suffer from over-fitting under the standard Gaussian ('*rbf*') kernel because of the complementary or near complementary sensitivities and specificities in the classifications. For instance, two algorithms both have 100% average sensitivity and 0% specificity on the ovarian data. It indicates that both algorithms can only detect samples from the majority type (cancer) on the ovarian data under the '*rbf*' kernel. Moreover, the average classification rates for the two algorithms are 64.13%, which are near to the majority type ratio for this dataset, i.e., $162/253=64.03\%$. 2). There is no over-fitting associated with the NPCA-SVM algorithm under the '*rbf*' kernel. Alternatively, it has achieved exceptional sensitivities and specificities for all datasets. 3). The PCA-SVM algorithm has achieved comparable results as the SVM algorithm under a linear kernel. This shows that global features chosen from PCA will not contribute to improving pattern prediction ratios. 4). The NPCA-SVM algorithm has obviously leading advantages over the two siblings. Its average specificities for the two ovarian datasets achieved 99%+ that is the population screening requirement ratio in general clinical diagnostics.

We further verify our algorithm's superiority by comparing it with other four peers, which are k-nearest neighbor (k-NN) [21], principal component analysis based linear discriminant analysis (PCA-LDA)[22], non-negative matrix factorization based support vector machine (NMF-SVM)[23], and independent component

analysis based support vector machine (ICA-SVM)[23]. For each dataset, we still use the same 100 trials of training and testing data from 50% HOCV in classification.

Table 2: Comparison of three algorithm performances

Dataset	$E(r_c)$	$E(r_s)$	$E(r_p)$
Ovarian			
<i>npca-svm-linear</i>	98.94±00.65	98.35±01.03	99.98±00.24
<i>npca-svm-rbf</i>	99.79±00.35	100.0±00.00	99.42±00.99
<i>svm-linear</i>	99.50±00.83	100.0±00.00	98.63±02.21
<i>svm-rbf</i>	64.13±02.88	100.0±00.00	00.00±00.00
<i>pca-svm-linear</i>	99.96±00.26	99.98±00.17	99.93±00.51
<i>pca-svm-rbf</i>	64.13±02.88	100.0±00.00	00.00±00.00
Ovarian-qaqc			
<i>npca-svm-linear</i>	98.70±00.89	98.01±01.94	99.27±00.90
<i>npca-svm-rbf</i>	98.91±00.98	98.11±02.25	99.57±00.82
<i>svm-linear</i>	96.57±01.99	96.16±03.52	96.97±02.19
<i>svm-rbf</i>	54.92±44.80	03.00±17.18	97.00±17.18
<i>pca-svm-linear</i>	97.12±01.17	97.14±02.16	97.94±01.57
<i>pca-svm-rbf</i>	54.95±44.70	03.20±17.22	96.80±17.22
Liver			
<i>npca-svm-linear</i>	96.02±01.35	97.68±01.71	94.40±02.22
<i>npca-svm-rbf</i>	97.25±01.30	98.35±01.67	96.20±02.01
<i>svm-linear</i>	91.78±02.27	91.04±03.76	91.04±03.76
<i>svm-rbf</i>	47.92±02.00	38.00±48.78	62.00±48.78
<i>pca-svm-linear</i>	90.21±01.99	90.96±03.69	89.57±03.56
<i>pca-svm-rbf</i>	47.92±02.00	38.00±48.78	62.00±48.78
Colorectal			
<i>npca-svm-linear</i>	98.14±01.27	97.93±02.32	98.35±02.00
<i>npca-svm-rbf</i>	97.15±01.07	95.81±02.78	98.18±02.22
<i>svm-linear</i>	96.55±01.87	94.35±03.47	98.26±02.16
<i>svm-rbf</i>	55.64±05.55	04.00±19.69	96.00±19.69
<i>pca-svm-linear</i>	93.21±03.38	92.59±04.68	93.89±05.56
<i>pca-svm-rbf</i>	56.25±05.27	03.45±17.50	97.25±15.60

The four comparison algorithms can be categorized into two types. The k-NN and PCA-LDA algorithms are widely used algorithms in proteomic pattern prediction. As a simple Bayesian inference method, k-NN determines an unknown sample class type according to the voting of its neighbor samples' class types. As a subspace classification method, PCA-LDA determines an unknown sample class type by employing linear discriminant analysis (LDA) in the subspace spanned by the principal components of the training data [22]. On the other hand, the NMF-SVM and ICA-SVM algorithms are feature selection based support vector machine algorithms. Similar to the NPCA-SVM algorithm, they conduct SVM classification for the meta-samples of an input dataset computed through nonnegative matrix factorization (NMF), and independent component analysis (ICA) respectively. Detailed information about the

three algorithms: LDA, NMF and ICA can be found in [24, 6, 25].

The four comparison algorithms have the following implementations. The distance measures are chosen as the correlation and Euclidean distances and the nearest neighbor number for an unknown sample is selected as $k = 2 \sim 7$ in k-NN. The average classification ratio achieved at the optimal value of k is counted as the final classification rate for k-NN among the 6 rounds of 100 trials of classifications. The matrix decomposition rank is selected as $r = 2 \sim 10$ in NMF-SVM. The best average classification rate among the 9 rounds of 100 trials of classifications is counted as the final average classification rate for the NMF-SVM algorithm. The number of independent components (ICs) is selected as the dimension of the input data in the ICA-SVM algorithm.

Table 3 shows the average average classification rates, sensitivities and specificities of these four algorithms and corresponding standard deviations for each dataset. We have the following observations from these results. 1). The k-NN algorithm achieves better classification performances under the correlation distance than the Euclidean distance. Similarly, the NMF-SVM and ICA-SVM algorithms both demonstrate classification advantages under the linear kernel over the standard Gaussian kernel. In addition, it seems that the NMF-SVM algorithm avoids over-fitting under the standard Gaussian kernel at cost of low sensitivities and specificities. However, it is interesting to see that the ICA-SVM algorithm still can not avoid this problem. 2). It seems that the NMF-SVM and k-NN algorithms have generally comparable classification performance on the four datasets. Similarly, the ICA-SVM and PCA-SVM algorithms generally have the same level of performance. It is also clear that they both outperform the k-NN and NMF-SVM algorithms. Although the PCA-LDA algorithm has achieved the best performance among all the six comparison peers, it still can not compete with the NPCA-SVM algorithm for all four profiles. Unlike the instabilities showed in the five comparison algorithm, our algorithm shows consistently leading performance for all datasets. The smaller standard deviations of the three classification measures in the NPCA-SVM algorithm also provide strong support for this observation.

Figure 1 compares our algorithm's performance with those of the four comparison peers: ICA-SVM, PCA-SVM, SVM and PCA-LDA one behalf of the average classification rates, sensitivities, specificities and negative target prediction ratios. Since there are over-fitting associated with the PCA-SVM, SVM and ICA-SVM classifications under the 'rbf' kernel, we only include their performance under the linear

kernel. It is obvious that the NPCA-SVM algorithm has demonstrated superior or comparable performance compared to the other four algorithms on all datasets under the 'rbf' and 'linear' kernels.

Table 3: Performance of the other four comparison algorithms

Dataset	$E(r_c)$	$E(r_s)$	$E(r_p)$
Ovarian			
<i>nmf-svm-linear</i>	97.41±00.94	99.91±00.31	92.92±02.50
<i>nmf-svm-rbf</i>	94.29±02.72	96.27±03.35	90.83±04.48
<i>knn-correlation</i>	96.53±01.57	99.28±01.34	91.67±03.67
<i>knn-euclidean</i>	96.41±01.29	99.58±00.76	90.77±03.19
<i>pca-lda</i>	99.67±00.87	99.93±00.38	99.21±02.00
<i>ica-svm-linear</i>	99.99±00.08	99.99±00.12	100.0±00.00
<i>ica-svm-rbf</i>	64.13±02.88	100.0±00.00	00.00±00.00
Ovarian-qaqc			
<i>nmf-svm-linear</i>	88.69±03.47	92.02±05.01	86.24±05.67
<i>nmf-svm-rbf</i>	77.30±03.67	76.18±09.12	78.57±06.38
<i>knn-correlation</i>	90.87±02.92	89.99±04.68	91.82±04.43
<i>knn-euclidean</i>	85.03±03.71	82.03±06.86	87.71±05.86
<i>pca-lda</i>	97.69±00.65	98.81±01.68	96.99±00.03
<i>ica-svm-linear</i>	97.56±01.45	97.80±02.46	97.41±01.77
<i>ica-svm-rbf</i>	62.61±07.11	20.98±17.05	96.76±11.75
Liver			
<i>nmf-svm-linear</i>	77.76±02.48	84.58±05.14	71.30±05.12
<i>nmf-svm-rbf</i>	74.79±02.25	80.69±06.01	69.21±05.57
<i>knn-correlation</i>	76.48±02.20	72.27±04.60	80.80±04.57
<i>knn-euclidean</i>	76.11±02.51	77.04±05.81	75.38±05.33
<i>pca-lda</i>	90.08±02.13	91.39±03.53	88.87±03.95
<i>ica-svm-linear</i>	86.61±02.87	87.78±04.55	86.50±04.86
<i>ica-svm-rbf</i>	60.24±06.42	56.51±29.90	66.80±28.54
Colorectal			
<i>nmf-svm-linear</i>	94.73±03.09	92.71±06.14	96.49±03.45
<i>nmf-svm-rbf</i>	93.61±02.88	91.21±05.97	95.65±04.40
<i>knn-correlation</i>	95.05±03.17	96.17±02.91	94.28±05.33
<i>knn-euclidean</i>	91.91±02.66	90.57±05.94	93.11±04.11
<i>pca-lda</i>	94.05±02.78	94.16±03.74	94.01±04.12
<i>ica-svm-linear</i>	96.04±02.02	94.38±03.66	97.39±02.97
<i>ica-svm-rbf</i>	60.04±07.99	12.58±16.91	97.94±13.70

4 Over-fitting Analysis.

Some important questions still remain to be answered, i.e., why did the SVM and PCA-SVM algorithms encounter the over-fitting problem at the standard Gaussian kernel? What made their final sensitivities and specificities complementary to each other? In this section, we answer these questions through a rigorous mathematical analysis.

Mathematically, the kernel matrix for a support vector machine contains all knowledge to categorize

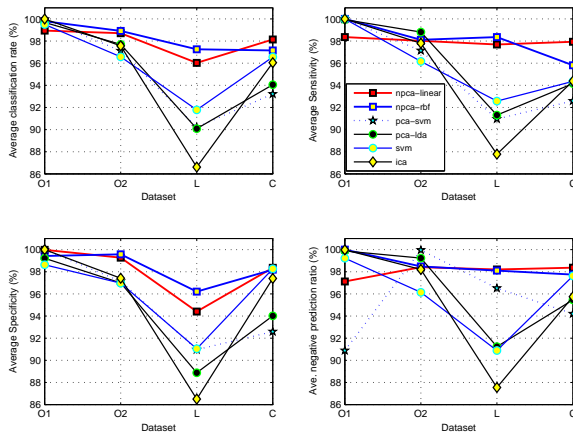


Figure 1: Comparison on the classification performance of four algorithms on four datasets: ‘O1’ (ovarian), ‘O2’ (ovarian-qaqc), ‘L’ (liver), ‘C’ (colorectal) The NPCA-SVM algorithm has demonstrated better or comparable performance.

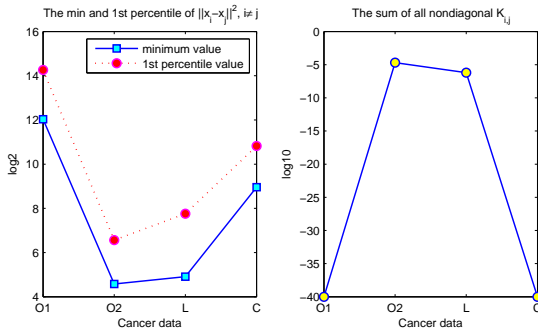


Figure 2: The minimum and 1st percentile of $\|x_i - x_j\|^2$ and the sum of non-diagonal entries of the kernel matrix for each dataset.

an input unknown sample. If the kernel matrix is an identity matrix or near the identity matrix that is a matrix with near-zero non-diagonal entries, it can only represent the concept of identity, i.e., it may perform very well for one or several training sets, but it does not have capability to generalize to other datasets. This leads to the over-fitting problem for the learning machine. Thus, it will be interesting to look at the kernel matrices generated under the standard Gaussian kernel for the mass spectral profiles. For convenience, we treat all samples in each mass spectral profile as training samples. i.e., the population data is viewed as the training dataset and it includes all possible training samples in a SVM implemented by any cross validations.

We have found the kernel matrices of the four protein profiles under the standard Gaussian kernel, i.e., $k(x, y) = e^{-\|x-y\|^2/2}$, where x, y represent two samples in

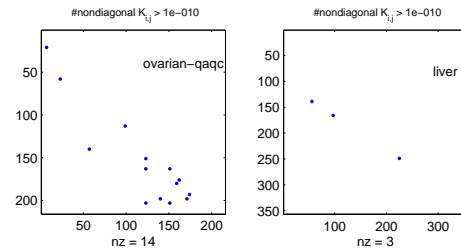


Figure 3: The SVM kernel matrix non-diagonal entries (‘rbf’) on the ovarian-qaqc and liver data. Only the lower triangle elements are visualized.

our context, are the identity or near identity matrices. The left figure in Figure 2 shows that the minimum values of all possible square distances between the samples. Exactly, they are $2^{12.0336}$ and $2^{8.9585}$ for the ovarian and colorectal data, and $2^{4.5777}$ and $2^{4.9092}$ for the ovarian-qaqc and liver data respectively. It is easy to find that any non-diagonal kernel entry is $e^{-2^{11}}$ for the ovarian data and $< e^{-2^{7.95}}$ for the liver data, i.e., their kernel matrices are the identity matrices. This point is also supported by the sums of all non-diagonal entries in their kernel matrices, i.e., $\sum_{i \neq j} k_{ij} \sim 0$ from the right figure in Figure 2.

Alternatively, we also find that the 1st percentile of all possible square distances between different samples in the ovarian-qaqc and liver datasets are $2^{6.5652}$ and $2^{7.7573}$ respectively. That is, their kernel matrix non-diagonal entries are $< e^{-5.56} (\sim 10^{-21})$ and $< e^{-6.75} (\sim 10^{-47})$ in a 99% confidence interval. Obviously, their kernel matrices will be near identity matrices. Correspondingly, the sums of their non-diagonal kernel entries are 2.026×10^{-5} and 6.3×10^{-6} respectively. Figure 3 shows that there are only 14 and 3 non-diagonal kernel entries $k_{ij} > 10^{-10}$ in the lower triangle parts of the ovarian-qaqc and liver data kernel matrices respectively.

According to these results, we say that the SVM kernel matrices under the standard Gaussian kernel for the four protein expression profiles will always be the identity matrices or near identity matrices, no matter which types of cross validations are employed in the implementation. To deepen our over-fitting analysis, theorem 4.1 rigorously proves that a support vector machine can only recognize majority type samples when its kernel matrix is the identity or near identity matrix.

THEOREM 4.1. *Let $X = [x_1, x_2 \dots x_d]^t$, $x \in \mathbb{R}^n$ be a training dataset with d samples across n pseudo-genes*

($d \ll n$) with labels $c = [c_1, c_2 \dots c_d]^t$, $c_i \in \{-1, 1\}$, drawn from a dataset $D \in \mathbb{R}^{N \times n}$ ($N > d$), input a standard SVM with a kernel function $k(x, y)$, then, for a testing sample $x' \in D - X$, if the kernel term is zero or approximately zero for a pair of samples: $k(x, x') \sim 0$, $\forall x, x' \in D, i \neq j$, its class type can be determined by the following decision function,

$$(4.9) \quad f(x') = \begin{cases} 1 & \text{if } |\{c_i | c_i = 1\}| > |\{c_i | c_i = -1\}| \\ 0 & \text{if } |\{c_i | c_i = 1\}| = |\{c_i | c_i = -1\}| \\ -1 & \text{if } |\{c_i | c_i = 1\}| < |\{c_i | c_i = -1\}| \end{cases}$$

According to theorem 4.1 and previous results in this section, we say that a support vector machine (SVM) with the standard Gaussian kernel may not avoid over-fitting for a mass spectral dataset, because of the large square distances $\|x_i - x_j\|^2$ between two samples, which directly leads to the identity or near identity kernel matrix. Moreover, the average classification rate under the 100 trials of 50% HOCV will generally approximate the majority type ratio in each dataset. For example, the average SVM classification rate for the ovarian data is 64.13% approximating the dataset majority type ratio $162/253=64.03\%$. All ‘local’ majority types in 100 sets of training data are just the majority type in the ovarian dataset. Thus, the average sensitivity and specificity are 100% and 0% respectively.

However, it is likely that the minority type of a dataset is sampled as a ‘local’ majority type in some trials. The average classification rate will be lower than the majority type ratio in the dataset. For example, the average SVM classification rate, sensitivity, and specificity for the liver data are 47.92%, 38.00%, and 62.00% respectively. It means that the minority type (‘control’) in the liver data is sampled as the ‘local majority type’ among 38 trials and there are only 62 trials where their local majority types are just the majority type (‘cancer’) of the dataset. Finally, the average classification accuracy 47.92% is less than the majority type ratio: $181/357=50.70\%$. In the 100 trials of 50% HOCV, the number of the minority type sampled as the ‘local majority type’ and the number of the majority type sampled as the ‘local majority type’ are always complementary to each other. This is the reason why the average sensitivities and specificities are complementary to each other for all the datasets in the 100 trials of 50% HOCV support vector machine classifications.

Similarly, we only need to prove that its kernel matrix is also an identity or near the identity matrix, to analyze over-fitting of the PCA-SVM learning machine at the standard Gaussian kernel. Theorem 4.2 states the structure of the kernel matrix in the PCA-SVM learning machine.

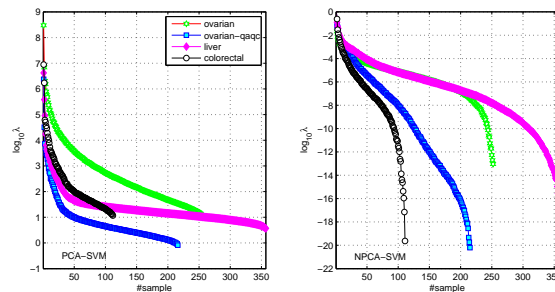


Figure 4: Corresponding lambda values of four protein expression datasets.

Table 4: Ranges of λ values and $e^{-(\lambda_i + \lambda_j)n/2}$

Dataset	$\log_{10} \lambda$	$e^{-(\lambda_i + \lambda_j)n/2}$
Ovarian	1.1233–8.4782	$\leq e^{-3360.6061}$
Ovarian-qaqc	-0.0875–6.3762	$\leq e^{-176.6949}$
Liver	0.5677–6.6245	$\leq e^{-1319.2597}$
Colorectal	1.0685–6.9465	$\leq e^{-1311.2541}$

THEOREM 4.2. Let $S = [x_1, x_2 \dots x_n]^t$, $x_i \in \mathbb{R}^m$, be a training dataset with n samples across m pseudo-genes for a PCA-SVM learning machine with a standard Gaussian kernel, then its kernel matrix K has the following entries: $k_{ij} = \exp(-\frac{\lambda_i + \lambda_j}{2} - \frac{(u_i - u_j)^t T (u_i - u_j)}{2})$ for $i \neq j$, and $k_{ij} = 1$ for $i = j$, where $T = \frac{1}{n} S^t \bar{I} S$ is a rank-one matrix (\bar{I} is a $n \times n$ matrix, each element of which is 1), and λ_i and λ_j are the data variance values on the i^{th} and j^{th} PCs of S : u_i and u_j .

We show the kernel matrices of the PCA-SVM learning machine (‘rbf’ kernel) are the identity or near identity matrices under the 50% HOCV. According to theorem 4.2, each non-diagonal entry in the kernel matrix is related to the sample number n and data variance term λ . The left sub-figure in Figure 4 shows that the λ values of four datasets are in the range $10^{-0.0875} \sim 10^{8.47}$. Since the training data sample size n is in the order 10^2 and $(\lambda_i + \lambda_j)/2$ is in the range of $10^{1.9125} \sim 10^{10.4782}$, the term $e^{-(\lambda_i + \lambda_j)/2}$ in the non-diagonal k_{ij} is generally zero under the 50% HOCV. Table 4 shows the range of the $\log_{10} \lambda$ and $e^{-(\lambda_i + \lambda_j)/2}$. It is clear to see that the kernel matrices are the identity or near identity matrices, i.e., the PCA-SVM learning machine also can not avoid over-fitting under the standard Gaussian kernel. Applying the results from theorem 4.1, we also know that an unknown sample type is always determined by the majority type in the training data. Finally, the average classification rate is near to the majority type ratio in the input

dataset and the average sensitivity and specificity are complementary to each other.

In summary, the over-fitting problems in the SVM and PCA-SVM learning machines are caused by the large data variance between biological samples, which is indicated by the lambda values in the mass spectral datasets. The large values of the variation terms $\|x_i - x_j\|^2$ and $\|Su_i - Su_j\|^2$ lead directly an identity or a near identity kernel matrix. Unlike the SVM and PCA-SVM algorithms, the NPCA-SVM algorithm inputs the nonnegative PC matrix U , which goes through normalization and sparse coding, into a support vector machine (SVM). The variances between biological samples are decreased by the nonnegativity constraints, sparse coding and normalizations. The nonnegativity constraints guarantee purely additive data representations, i.e., global features are decomposed as simple additions of local features. Correspondingly, the large variances from the global features are decomposed as additions of small variances of local features. The sparse coding brings more zeros in the nonnegative principal components and decreases the total data variances. The normalization scales all PC matrix entries to uniformly small values, which also contributes to a decrease in data variances.

The right sub-figure in Figure 4 shows that the λ values for the four protein expression datasets in the range $10^{-20.2014} \sim 10^{0.0}$. It indicates the small variances between biological samples in the NPCA-SVM algorithm. Figure 5 shows the minimum, median and sum of the non-diagonal kernel entries for the four datasets in the NPCA-SVM algorithm in the left and right sub-figures respectively. Just as before, the whole population data is treated as the training data. Clearly, the kernel matrix in the NPCA-SVM classification for each dataset is not the identity or near identity matrix, i.e., there is no over-fitting associated with the nonnegative principal component based support vector machine algorithm.

What is the biological reason for over-fitting associated with the ‘*rbf*’ kernel for the PCA-SVM, SVM and ICA-SVM algorithms? The reason is from the sensitive signal-amplification mechanism from the serum proteomics technology itself. In serum proteomics, almost any small even tiny changes in the part of proteome will be amplified to rather large even huge differences in mass spectra, no matter the sources of the changes are from biological factors or experimental conditions. The exponential transform in the ‘*rbf*’ kernel makes the already amplified expression values of two biological samples has zero or almost zero distance in the corresponding SVM kernel space. In other words, the SVM learning machine inevitably loses its detection

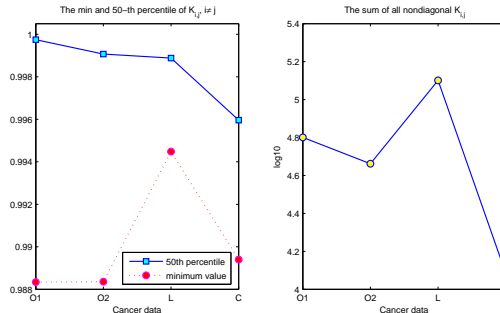


Figure 5: The minimum, 50th percentile, and sum of non-diagonal entries of the kernel matrix for each profile.

capabilities.

5 Biomarker Discovery by Nonnegative Principal Component Analysis.

Nonnegative principal component analysis can be also applied to capture biomarkers for protein expression data. We present a nonnegative principal component analysis based filter-wrapper biomarker discovery algorithm in this section and apply it to the colorectal and ovarian data. The Bayesian two-sample t-test [26, 27] and nonnegative principal component analysis function as filters and a support vector machine works as a wrapper. Unlike other biomarker capturing methods in proteomics [17, 18], our algorithm can identify which pseudo-genes are more effective in predicting cancer patterns. The NPCA-based biomarker discovery approach can be described as follows.

For an input mass spectral data $X \in \mathbb{R}^{n \times m}$ with m pseudo-genes and n biological samples, we first filter a potential biomarker set by conducting the two-sample Bayesian t-test, which is a novel approach to evaluate pseudo-genes according to their differentially expressed levels. The potential biomarker set S_b consists of significantly differentially-expressed pseudo-genes. For each dataset, we select at least the top 1% pseudo-genes with the smallest Bayesian factors, i.e., $|S_b| = \lceil m \times 0.01 \rceil$, to construct S_b .

Then, nonnegative principal component analysis is employed to decompose the input data, i.e., $X^t \sim PU^t$. For each pseudo-gene, a coefficient τ is used to rank its contribution to all PCs. For example, the coefficient for the i^{th} pseudo-gene is calculated as the weighted sum of the i^{th} row in the nonnegative matrix P : $\tau_i = \log \sum_{j=1}^{\#PC} w_j p_{ij}$, where $w_j = \lambda_j / \sum_{i=1}^{\#PC} \lambda_i$ is the ratio of variance explained in j^{th} the PC among the total data variance. A large coefficient value of a pseudo-gene indicates it has significant contributions to the PCs.

Each pseudo-gene in S_b is used to train a SVM classifier under the leave-one-out cross validation

Table 5: Biomarkers captured for the colorectal data

m/z	Bayes factor	nPCA-coefficient	SVM ratio
969.1849	7.7881e-031	-1.1205	0.9643
997.5336	1.4236e-026	-1.1571	0.9018
1016.389	7.6644e-013	1.2773	0.8152

(LOOCV). The first biomarker g_1 is selected as the pseudo-gene with the highest accuracy. If there is more than one candidate, the pseudo-gene with the largest coefficient in NPCA-ranking will be selected. The potential biomarker set is updated by removing the selected biomarker: $S_b = S_b - \{g_1\}$. The second biomarker pseudo-gene g_2 is selected from the current potential-biomarker set S_b such that the SVM classifier reaches its maximum classification rate for the combination of g_1 and g_2 . If there is more than one candidate, the pseudo-gene with the largest coefficient in the NPCA-ranking will be chosen as g_2 . Similarly, the set S_b is further updated as $S_b = S_b - \{g_2\}$. Such a proceeding continues until the SVM classifier achieves the maximum classification accuracy with the fewest biomarkers.

We apply the nonnegative principal component analysis based biomarker capturing algorithm to the colorectal dataset. The potential biomarker set S_b is initialized by 200 pseudo-genes with the smallest Bayes factors. The alpha value in NPCA is set as $\alpha = 10$ to maintain consistency with the previous classification. Table 5 shows the information about three biomarkers discovered for the colorectal data. The SVM ratio for a biomarker is the classification ratio of the SVM classifier under this pseudo-gene in the leave-one-out-cross-validation (LOOCV) under a linear kernel. The total SVM accuracy under the three biomarkers is 98.21% and the corresponding sensitivity and specificity are 95.83% and 100% respectively.

It is interesting that these biomarkers are the peaks not with very large intensity values. The similar results can also be obtained by running the biomarker capturing algorithm under the ‘*rbf*’ kernel. The final SVM accuracy also reaches 98.21% with three biomarkers at the 970.0379, 973.1689 and 997.5336 m/z ratios. It is also interesting to find that the biomarkers from different kernels not only share a same pseudo-gene (at 997.5336 m/z), but also show a spatial coherence, i.e., they are neighbors close or very close to each other among 16331 m/z ratios in the data. It indicates that m/z ratios in the downstream interval (960, 1030) may be more sensitive in discovering cancer patterns than others. Figure 6 visualizes all samples of the colorectal data by using three biomarkers found under the linear kernel. It is clear to see that the 112 samples consist of two parts:

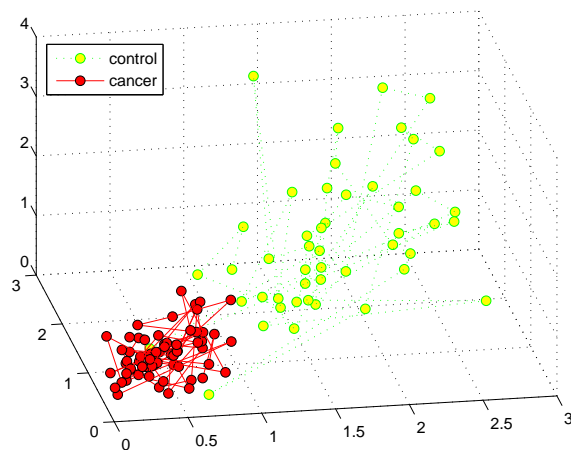


Figure 6: Visualization of 112 colorectal samples by using three biomarkers

64 cancers and 48 controls though one control sample is ‘wired’ in the region of cancer samples.

Similarly, we also applied this algorithm to the ovarian data and obtained 100% prediction accuracy (sensitivity: 100%, specificity: 100%) from four biomarkers at m/z ratios: (0.452124, 0.000096, 0.530561, 1.276201) under the linear kernel. Moreover, The SVM classifier also achieved 99.60% accuracy, 100% sensitivity, and 98.90% specificity under the ‘*rbf*’ kernel from three biomarkers at m/z ratios: (0.464762, 0.000096, 0.517053). Also similar to the previous case, the biomarkers discovered under different kernels illustrated spatial proximity and shared same pseudo-genes. Figure 7 visualizes all 253 samples by using the three biomarkers obtained from the ‘*rbf*’ kernel. It is interesting to see that cancer and control samples are separated into two disjoint clusters.

6 Discussion and Conclusion.

In this work, we developed a novel feature selection algorithm: nonnegative principal component analysis, and proposed the nonnegative principal component analysis based support vector machine under the sparse coding for a high performance proteomic pattern classification. We demonstrate the superiority of our algorithm by comparing it with other six peer algorithms on four mass spectral profiles. Moreover, we have mathematically proved that the over-fitting problem is unavoidable for the SVM and PCA-SVM algorithms under the standard Gaussian kernel on a mass spectral dataset. However, the over-fitting problem can be overcome by the NPCA-SVM algorithm with exceptional classification

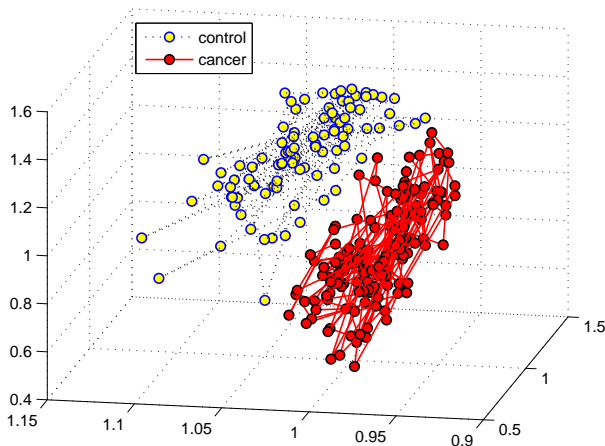


Figure 7: Visualization of 253 samples in the ovarian data by using three biomarkers

performance. In addition, we have designed a NPCA-based filter-wrapper biomarker capture algorithm and applied it to effectively capture meaningful biomarkers for the colorectal and ovarian data.

Although nonnegative principal component analysis has overcome the global nature of the standard PCA algorithm and contributed to the high-performance proteomic pattern prediction and effective biomarker discovery, it is an expensive algorithm with a high complexity $O(d^2n \times N)$ compared to the classic PCA algorithm ($O(d^3)$) for an input data $X \in \mathbb{R}^{d \times n}$. It may require some basic feature selection preprocessing such as the two-sample t-test to avoid a large computing burden for a high-dimensional dataset. On the other hand, since the final PC matrix in nonnegative principal component analysis is computed through a fixed instead of an optimal step size in the iteration, it may miss some local optimal solutions and lead to potential convergence problems.

Our algorithm developed could be extended to a multi-resolution nonnegative principal component analysis by employing a discrete wavelet transform to overcome the NPCA's high algorithm complexities, where a discrete wavelet transform is employed to decompose an input data into a multi-resolution form. The nonnegative principal component analysis (NPCA) is only conducted in the fine level wavelet coefficients to extract the local data features and suppress wavelet coefficients in the coarse level because they generally represent the global features of data. In addition, we will employ a projected-gradient algorithm with a dynamic step size to improve the nonnegative principal component analysis algorithm convergence [28].

As a local feature selection algorithm, nonnegative principal component analysis can be integrated with other state-of-the-art learning machines to develop a family of statistical learning algorithms. For instance, we are interested in investigating combining it with the linear programming SVM algorithm [29] to further explore its potentials in mass spectra prediction. In addition, we will continue to investigate the applications of the NPCA-SVM algorithms in the SNP [30] and CGH array data [31], and other related topics [32, 33, 34] in future work.

References

- [1] K. R. Coombes, J. Morris, J Hu, S. Edmonson, and K. Baggerly K, *Serum proteomics profiling – a young technology begins to mature*, Nat. Biotechnol., 23 (2005), pp. 291–292.
- [2] H. Hauskrecht et al., *Feature Selection for Classification of SELDI-TOF-MS Proteomic Profiles*, Applied Bioinformatics, 4, (2005), pp. 227–246.
- [3] J. S. Yu, S. Ongarello, R. Fiedler, X. W. Chen, G. Toffolo, C. Cobelli, and Z. Trajanoski, *Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data*, Bioinformatics, 21, (2005), pp. 2200–2209.
- [4] D. Mantini, et al, *Independent component analysis for the extraction of reliable protein signal profiles from maldi-tof mass spectra*, Bioinformatics, 24(1), (2008), pp. :63–70.
- [5] D. Lee and H. Seung, *Learning the parts of objects by non-negative matrix factorization*, Nature, 401, (1999), pp. 788–791.
- [6] P. Hoyer, *Non-negativematrix factorization with sparseness constraints*, Journal of Machine Learning Research 5, (2004), pp. 1457–1469.
- [7] K. Noy and D. Fasulo, *Improved model-based, platform-independent feature extraction for mass spectrometry*, Bioinformatics, 23(19), (2007), pp. 2528–2535.
- [8] I.T. Jolliffe, *Principal component analysis*, Springer Series in Statistics, 2nd ed., Springer, New York, 2002.
- [9] C. Cortes, and V.N. Vapnik, *Support vector networks*, Machine Learning, 20, (1995), pp. 273–297
- [10] B. Schölkopf, *Support Vector Learning*, R. Oldenbourg Verlag, Munich, 1997.
- [11] V.N. Vapnik, *Statistical Learning Theory*, John Wiley & Son, Inc., New York, 1998.
- [12] N. Cristianini, N. and J. Shawe-Taylor, *An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods)*, Cambridge University Press, Cambridge, 2000.
- [13] R. Zass and A. Shashua, *Nonnegative sparse PCA*, Neural Information and Processing Systems (NIPS), 2006.
- [14] J. Nocedal, and S. Wright, *Numerical Optimization*, Springer, New York, 1999.

- [15] S. Boyd, and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.
- [16] NCI Cancer Proteomics Program: <http://home.ccr.cancer.gov/ncfdaproteomics>
- [17] H. Resson et al, *Analysis of mass spectral serum profiles for biomarker selection*, , Bioinformatics 21(21), (2005), pp. 4039–4045.
- [18] T. Alexandrov et al, *Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation*, Bioinformatics, 25(5), (2009), pp. 643–649.
- [19] A. Cruz-Marcelo et al, *Comparison of algorithms for pre-processing of SELDI-TOF mass spectrometry data*, Bioinformatics, 24(19), (2008), pp. 2129–2136.
- [20] P. Du et al, *Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching*, Bioinformatics 22, (2006), pp. 2059–2065.
- [21] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2nd edition, Springer, New York, 2009.
- [22] R. Lilien, and H. Farid, *Probabilistic Disease Classification of Expression-dependent Proteomic Data from Mass Spectrometry of Human Serum*, Journal of Computational Biology, 10 (6), (2003), pp. 925–946.
- [23] X. Han, *Nonnegative Principal Component Analysis for Cancer Molecular Pattern Discovery*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, IEEE computer Society Digital Library, 2009.
- [24] A. Martinez, et al, *PCA versus LDA*, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2), (2001), pp. 228–233.
- [25] A. Hyvärinen, *Fast and robust fixed-point algorithms for independent component analysis*, IEEE Transactions on Neural Networks, 10(3), (1999), pp. 626–634.
- [26] M. Gonen, W. Johnson, Y. Lu and P. Westfall, *The Bayesian Two-Sample t Test*, The American Statistician, 59(3), (2005), pp. 252–257.
- [27] R. Fox and M. Dimmic, *A two-sample Bayesian t-test for microarray data*, BMC Bioinformatics, 7(126), (2006).
- [28] C. Lin, *Projected gradient methods for non-negative matrix factorization*, Neural Computation. 19(10), (2007), pp. 2756–2779.
- [29] A. Chan, et al, *Direct Convex Relaxations of Sparse SVM*, Proceedings of the 24th International Conference on Machine Learning, 2007.
- [30] C. Li et al, *Major copy proportion analysis of tumor samples using SNP arrays*, BMC Bioinformatics, (2008).
- [31] J. Liu, S. Ranka and T. Kahveci, *Classification and feature selection algorithms for multi-class CGH data*, Bioinformatics, 24 ISMB, (2008), pp. i86–i95.
- [32] X. Li, Y. Tan and S. Ng, *Systematic gene function prediction from gene expression data by using a fuzzy nearest-cluster method*, BMC Bioinformatics, 7(Suppl 4): S23, (2006).
- [33] X. Chen, J. Jeong, *Sequence-based prediction of protein interaction sites with an integrative method*, Bioinformatics 25 (5), (2009), pp. 585–591.
- [34] M. Kirac, G. Özsoyoglu and J. Yang, *Annotating proteins by mining protein interaction networks*, , ISMB (Supplement of Bioinformatics), (2006), pp. 260–270.