Original Research Article

On Designing Invertible Pseudo Covariance Matrix for Undersampled Cases in Classification

ABSTRACT

In linear discriminant analysis, determinant and inverse of the covariance matrix are required to be computed. If number of features is greater than the number of available examples, covariance matrix is no longer invertible. A common approach is to reduce dimensionality due to which some features of interest may be lost. When we are not interested in dimensionality reduction, one approach to solve such problems is to take pseudoinverse of covariance matrix which is not always possible. We propose, in such cases, to project covariance matrix onto a highly correlated space to compute pseudoinverse of the matrix. Proposed solution has been tested for classification of microarray gene expression data of colon's tumor.

Keywords: Covariance matrix, dimensionality reduction, pseudoinverse

1. INTRODUCTION

Let x_i be data points or collection of observations such that $x_i \in \mathbb{R}^a$ for $1 \leq i \leq n$ and $i \in \mathbb{N}$, where $\mathbb{R}$ is a set of real numbers and number of data points be n. Therefore, $\mathbb{R}^{a \times n}$ is our sample space. In classification, a pattern is a pair of variables $\{x_i, w_i\}$ where w_i is the concept behind the observation. The task is to discriminate examples from different classes with respect to w_i. Examples from the same class should have similar w_i while examples from different classes have different w_i. The minimum-error-rate classification can be achieved by use of following discriminant functions:

$$g_i(\mathrm{x}) = lnp(\mathrm{x}|w_i) + \ln P(w_i) \qquad (1)$$

If $p(\mathrm{x}|w_i) \sim \mathcal{N}(\mu_i, \Sigma_i)$, we have

$$g_i(\mathrm{x}) = -\frac{1}{2}(\mathrm{x} - \mu_i)^t \Sigma_i^{-1}(\mathrm{x} - \mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i) \qquad (2)$$

The covariance matrix Σ is always symmetric and positive semi-definite. In the following, we discuss the general case when Σ_i is arbitrary.

In a general multivariate normal case, covariance matrices are different for each category. The only term that can be dropped from Eq. 2 is the $\frac{n}{2}\ln(2\pi)$ term, and the resulting discriminant functions are inherently quadratic.

$$g_i(\mathrm{x}) = \mathrm{x}^t W_i \mathrm{x} + \boldsymbol{w}_i{}^t \mathrm{x} + w_{i0} \qquad (3)$$

Where

$$W_i = -\frac{1}{2}\Sigma_i^{-1}$$

$$w_i = \Sigma_i^{-1}\mu_i$$

and

$$w_{i0} = -\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

This shows that computing inverse of covariance matrix (i.e., Σ^{-1}) is essential in discriminant analysis. In this situation, two types of problems arise when relationship between n and d is considered.

- Case 1 $(n \geq d)$: When number of available samples (n-samples) is equal to or greater than the number of dimensions (d-dimensions), the computation of Σ^{-1} is trivial.
- Case 2 $(n < d)$: In most of practical problems (e.g., classification of cardic signal or detection of tumor from microarray gene expression data), d is large such that $d \gg n$. Such problems are known as undersampled problems (Krzanowski et al., 1995).

There are three common approaches to solve these problems when n is smaller than d.

- Dimensionality reduction is applied such that $d < n$, where d is number of dimensions in reduced space. However, it is possible that some useful features may lose.
- Eigen values are translated such as none of eigen values is zero. This reshapes the system matrix.
- Computing pseudoinverse of Σ which is not always possible.

In this paper, we propose a solution to problem when Σ does not exist and dimensionality reduction and reshaping the system matrix are to be avoided. The related work is presented in Section 2. The proposed methodology is described in Section 3. Results and discussion are presented in Section 4. Finally, the paper is concluded is Section 5.

2. Related work

There are many algorithms developed so for to tackle undersampled problems (Akguna, 2016; Friedman, 1986; Yang, 2001; Yuen, 2003; Bensmail, 1995; Baudat, 2000; Raudys, 1998; Hastie, 1995; Howlan, 2004; Belhumeur, 1997). Most of the methods are proposed in the context of scatter matrix in classical linear discriminant analysis (LDA). This is because to perform discriminant analysis, the scatter matrix must be non-singular. Depending on nature of pre-processing of scatter matrix, the proposed methods can broadly be categorized in two categories: dimensionality reduction (DR) and matrix transformation (MT). The DR techniques are applied to reduce the dimensionality such that the features vectors have dimensions smaller than the number of samples. For instance in (Belhumeur, 1997; Swets, 1996; Zhao,

1999), principle component analysis (PCA) is used and orthogonalization is applied in (Zheng, 2004a; Zheng, 2004b) for reducing the dimensionality. Afterwards, LDA is applied for classification. DR is useful when we have to classify over the dominant features. However, in cases like classification of ventricle diseases using electrocardiogram waveforms or classification of cancerous genes using microarray data, classification may not be carried out using dominant features. Therefore, DR may not be helpful. In this situation, MT techniques can be used to make the scatter matrix invertible. The most notable examples of these techniques include regularized LDA (RLDA) (Friedman, 1986; Sullivan, 1986; Titterington, 1981), penalized LDA (PLDA) (Hastie et al., 1995), pseudo-inverse LDA (Raudys and Duin, 1998) and shrinkage LDA (SLDA) (Cornfield, 1967). In RLDA, all the eigen values of the scatter matrix are given a scalar, constant and equal shift. This replaces zero eigen values with some positive real value. This shift in eigen values makes the scatter matrix invertible (Friedman, 1986). This regularization minimizes within-class variance at the cost of bias. Therefore, we have to play with parameters(s) for optimal tradeoff between variance and bias. In penalized LDA, all values of within class scatter matrix are given a positive scalar shift, i.e., we use $\Sigma + \Omega$ where Σ is a within class scatter matrix and Ω the shift matrix. If $\Omega = \mu I$ for some constant μ and I as identity matrix, penalized LDA converges to the regularized LDA. Therefore, penalized LDA is more general as compared to regularized LDA. A problem with this technique is that there is no single algorithm (Ω) for any set of applications (Hastie, 1995). When the scatter matrix is singular, sometimes its pseudo-inverse is possible. Therefore, we consider pseudo-inverse of the scatter matrix in place of its inverse. The technique is referred as pseudo-inverse LDA. However, pseudo-inverse is not possible when numerical values are very small during computations. Pseudo-inverse is an estimate of the inverse of the matrix using least square estimation (Fukunaga, 1990; Raudys and Duin, 1998). Effect of the sample size and dimension on classification error was studied in (Krzanowski, 1995). Performance of this technique is comparable to PCA + LDA and RLDA (Skurichina and Duin, 1996).

3. METHODOLOGY

Let $\Sigma_b \in \mathbb{R}^{N \times N}$ be a space onto which Σ is projected. We consider that Σ_b is an ideal covariance matrix which is drifted by the covariance matrix of the system, i.e., Σ. We define Σ such that it is a projection of Σ on Σ_b.

$$\tilde{\Sigma} = \Sigma \Sigma_b \qquad (4)$$

Matrix Σ represents a situation (space) between ideal space and real space and possesses all the properties of Σ, like symmetry, Hermitian etc. The matrix Σ_b is defined such that elements along the main diagonal are *α* and all other elements are *β*; and *α* and *β* are real numbers. Some properties of Σ_b are given below.

3.1. Inverse

Let Σ_b be a matrix of size $n \times n$, as it is a symmetric Hermitian matrix, its inverse is computed as follows: Consider that μ_d, γ_d, $d_d \in R$, where d is the dimensionality and $d \in D$. Usually, matrix inversion is a computationally heavy task; we have derived simple expressions for its computation.

$$\mu_d = -\beta(\alpha - \beta)^{d-2} \tag{5}$$

$$\gamma_d = (\alpha - \beta)^{d-1} - (d-1)\mu_d \tag{6}$$

$$\tau_d = \mu_d + \beta\ (d-1)\mu_d \tag{7}$$

This τ_d is the determinant of Σ_b.

$$\Sigma_b^{-1} = \frac{1}{\tau_d}[\mu_d O_d + (\gamma_d - \mu_d) I_d] \tag{8}$$

where O_d is defined as a matrix of size $d \times d$ with all elements as *1*, and I_d is an identity matrix of the same size.

3.2. Eigenvalues of Σ_b

We have found that the largest eigenvalue of Σ_b is:

$$\lambda_{max} = \alpha + (d-1)\beta \tag{9}$$

Other eigenvalues are:

$$\lambda_i = \alpha - \beta \tag{10}$$

for $i = 2,3,\ldots,D$.

3.3. Eigenvalues of Σ_b^{-1}

As eigenvalues of Σ_b^{-1} are inverse of eigenvalues of Σ_b. Therefore, minimum eigenvalue is:

$$\tilde{\lambda}_{min} = \frac{1}{\alpha + (d-1)\beta} \tag{11}$$

and other $(d-1)$ eigenvalues are:

$$\tilde{\lambda}_i = \frac{1}{\alpha - \beta} \tag{12}$$

for $i = 2,3,\ldots,D$.

3.4. Eigen Value Decomposition of Σ_b

Consider that:

$$\Sigma_b = U_b \Lambda_b V_b^T \tag{13}$$

Λ_b is given by:

$$\Lambda_b = \begin{pmatrix} \alpha + (d-1)\beta & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha - \beta \end{pmatrix} \tag{14}$$

We have found U_b as:

$$U_b = \begin{pmatrix} -\frac{1}{\Delta_1} & -\frac{1}{\Delta_2} & \cdots & 0 & 0 \\ -\frac{1}{\Delta_1} & -\frac{1}{(d-1)\Delta_2} & & 0 & 0 \\ & \vdots & \ddots & \vdots & \\ -\frac{1}{\Delta_1} & \frac{1}{(d-1)\Delta_2} & \cdots & -\frac{1}{\Delta_{d-1}} & 0 \\ -\frac{1}{\Delta_1} & \frac{1}{(d-1)\Delta_2} & & -\frac{1}{2\Delta_{d-1}} & \frac{1}{\Delta_d} \end{pmatrix} \tag{15}$$

$\Delta_1 = \sqrt{d}$

and

$$\Delta_k = \sqrt{1 + \frac{1}{d-k+1}} \tag{16}$$

for $k = 2,3,\ldots,D$.

First column of U_b is an eigenvector corresponding to the largest eigenvalue of Σ_b. Matrix V_b is computed as:

$$V_b = U_b \tag{17}$$

3.5. Eigen Value Decomposition of Σ_b^{-1}

$$\Sigma_b^{-1} = (U_b \Lambda_b V_b^T)^{-1} \tag{18}$$

$$= U_b \Lambda_b{}^{-1} V_b^T \qquad (19)$$

The matrix $\Lambda_b{}^{-1}$ is a diagonal matrix with eigenvalues of Σ_b^{-1} on its diagonal.

3.6. Pseudo Inverse of $\hat{\Sigma}$

Let us consider that

$$\Sigma = U \Lambda V^T \qquad (20)$$

$$\Sigma_b = U_b \Lambda_b V_b^T \qquad (21)$$

And

$$\Sigma = \Sigma \Sigma_b$$

$$= U(\Lambda V^T U_b \Lambda_b) V_b^T \qquad (22)$$

$$= U \hat{\Lambda} V_b^T \qquad (23)$$

$$\hat{\Lambda} = \Lambda V^T U_b \Lambda_b \qquad (24)$$

We have found that under the assumption of $U = U_b$ and $V = V_b$:

$$V^T U_b = I \qquad (25)$$

$$\hat{\Lambda} = \Lambda \Lambda_b \qquad (26)$$

$$\hat{\Lambda}^{-1} = \Lambda_b{}^{-1} \Lambda^{-1} \qquad (27)$$

Therefore

$$\hat{\Sigma}^{\dagger} = U_b \hat{\Lambda}^{-1} V_b^T \qquad (28)$$

Where $\Sigma^{\dagger}$ represents the pseudoinverse of Σ.

3.7. Density Function in the Projected Space

The general expression for multivariate normal density in d-dimensions, in the projected space is as follows:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\hat{\Sigma}|^{\frac{1}{2}}} e^{[-\frac{1}{2}(\mathbf{x}-\mu)^t \hat{\Sigma}^{\dagger} (\mathbf{x}-\mu)]} \qquad (29)$$

Or simply $p(\mathbf{x}) \sim \mathcal{N}(\mu, \Sigma)$. The discriminant function corresponding to Eq. 29 is:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^T\hat{\Sigma}_i^{\dagger}(\mathbf{x}-\mu_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\hat{\Sigma}_i| + \ln P(\omega_i) \tag{30}$$

where Σ_i is arbitrary. We assume that Σ_i is also arbitrary. Therefore,

$$g_i(\mathbf{x}) = \mathbf{x}^t W_i \mathbf{x} + w_i^t \mathbf{x} + \omega_{i0} \tag{31}$$

Where

$$W_i = -\frac{1}{2}\hat{\Sigma}_i^{\dagger} \tag{32}$$

$$w_i = \hat{\Sigma}_i^{\dagger}\mu_i \tag{33}$$

And

$$\omega_{i0} = -\frac{1}{2}\mu_i^t\hat{\Sigma}_i^{\dagger}\mu_i - \frac{1}{2}\ln|\hat{\Sigma}_i| + \ln P(\omega_i) \tag{34}$$

3.8. Classification

Let $x_1^i, x_2^i, \ldots, x_N^i \sim (\mu_i, \Sigma)$ be data points from class i. Let d be the length of x_j, for $i \leq j \leq N$.

We compute the mean of the class.

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} x_j^i \tag{35}$$

Transforming $\mathcal{N}(\mu_i, \Sigma_i)$ to $\mathcal{N}(0, \Sigma_i)$:

$$x_j^i \leftarrow x_j^i - \frac{1}{d}\sum_{k=1}^{d} x_j^i[k] \qquad \forall j = 1, 2, \ldots, N$$

Where $x_j^i[k]$ means k^{th} value of jth sample in i^{th}class.

Let

$$X_i = [x_1^i\, x_2^i \ldots x_N^i]$$

The scatter matrix is:

$$\Sigma_i = \frac{1}{N}XX^T \tag{36}$$

$$\hat{\Lambda}_i = \Lambda_i\Lambda_i^b \tag{37}$$

Pseudo inverse is computed as:

$$\hat{\Sigma}_i^{\dagger} = U_b \hat{\Lambda}_i^{-1} U_b^T \qquad (38)$$

The classification algorithm produces μ_i and $\Sigma_i^{\dagger}$.

3.9. Threshold Setting

Eq. 31 is quadratic with d variables where d is large. Therefore, it is very hard to compute the closed form solution or any general expression for the threshold. We suggest that training samples, i.e., x_1^i, x_2^i, … for $i = 1, 2, \ldots, c$ are used in $g_i(x)$ given by Eq. 31 and human or machine intelligence is used for deciding threshold values, as optimal as possible.

3.10. Testing

Let y be a vector from unknown class. We transform distribution from y from $\mathcal{N}(\mu, \Sigma)$ to $\mathcal{N}(0, \Sigma)$ using:

$$y \leftarrow y - \frac{1}{d}\sum_{j=1}^{n} y[j]$$

where $y[j]$ is the jth value of the vector y. Compute $g_i(y)$ for $i = 1, 2, \ldots, c$ where c is the number of classes. Considering the threshold and $g_i(y)$, we decide the class to which y probably belongs.

4. RESULTS AND DISCUSSION

To analyze the effectiveness of proposed approach, we have taken a cancer gene expression dataset, namely “Colon Tumor”, which is collected from Kent Ridge Bio-medical Dataset Repository (Ridge). The dataset contains 62 instances collected from colon-cancer patients. There are 40 tumor biopsies from tumors. Other instances are normal biopsies taken from healthy parts of the same patients. Every sample has expressions from large number of genes and a class label (i.e. normal or cancer). The data has higher number of dimensions as compared to the number of samples, thus making the problem as low sample size problem, i.e., $\mathbf{d} \gg \mathbf{n}$, where d is dimension and n is the number of samples in the set. In this case, $d = 256$ and $n = 40$ and $n = 22$ for biopsies from sick and normal cells, respectively. As microarray experiments are very expensive, it seems impossible to have $d \leq n$.

This work does not include preprocessing of the raw dataset. Every sample is in the vector form. Such a form of data is the input to the method.

In order to test the effectiveness of our classifier, standard performance metrics are used in this research. Given a test set with N samples, let N_P and N_N be the number of positive samples ("normal") and the number of negative samples ("diseased") within the dataset ($N = N_P + N_N$) respectively. After classification, let T_P and F_P be the number of positives detected as being positive and the number of positives detected/classified as being negatives ($N_P = T_P + F_P$) respectively. Similarly, let T_N and F_N be

negatives detected as being negative and the number of negatives detected/classified as being positive ($N_N = T_N + F_N$). For this paper, we have considered the following metrics to analyze the performance.

1. Sensitivity: $Sn = TP/(TP+FN)$

2. Specificity: $Sp = TN/(TN+FP)$

3. Positive predicted value: $PPV = TP/(TP+FP)$

4. Negative predicted value: $NPV = TN/(TN+FN)$

5. Precision: $P = (TP+TN)/N$

The computed results are shown in the table 1.

Table 1. Performance analysis of the classifier.

Metric	Measure (%)
S_n	**66.67%**
S_p	**83.33%**
PPV	**75.00%**
NPV	**76.92%**
P	**76.19%**

5. CONCLUSION

We have studied a case in classification when dimensionality is larger than the number of samples. This paper discussed when neither the number of dimensions is reduced nor pseudoinverse is possible. We have proposed that in such cases projection of the scatter matrix onto an ideal scatter matrix makes computation of pseudoinverse possible. However, due to the computation of large power (Equations 5 to 8), the accuracy of the method can be affected. Moreover, the computational complexity and space complexity of the proposed technique is $O(d^3)$ where d the dimension is and d is very large. Therefore, it is a computationally heavy technique and requires large space, but the same is true for the existing techniques. The proposed classifier is also noise sensitive like existing techniques. However, by gaining a larger dataset and preprocessing, the results may further be improved.

REFERENCES

1. Akguna, Fatma Aydin, and Billy E. Rhoadesb. Characterizations of HJ Matrices. Filomat 30.3 (2016): 675-679.

2. Belhumeur, P. N. , Hespanha, J. P., Kriegman, D. J., 1997. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell., 19(7), 711-720.

3. Cornfield, J.,1967. Discriminant functions," Review of the International Statistical Institute, 35, 142-153.

4. Fukunaga, K., 1990. Introduction to statistical pattern classification. Academic press.

5. Hastie, T. et al., 1995. Penalized discriminant analysis. Ann. Stat., 73-102.

6. Howland, P., Jeon, M., Park, H., 2003. Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. Siam J. Matrix Anal. A.,, 25(1), 165-179.

7. Howland, P., Park, H., 2004. Generalizing discriminant analysis using the generalized singular value decomposition. IEEE Trans. Pattern Anal. Mach. Intell., 26(8), 995-1006.

8. Krzanowski, W. J. et al., 1995. Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. Appl. Stat-J. Roy St. C., 44, 101–115.

9. Raudys, S., Duin, R. P., 1998. Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. Pattern Recogn. Lett., 19(5), 385-392.

10. Ridge, K.,.2015. Kent ridge bio-medical dataset repository. Kent Ridge Bio-medical Dataset Repository.

11. Skurichina, M., Duin, R. P., 1996. Stabilizing classifiers for very small sample sizes. In proceedings of the 13th IEEE International Conference on in Pattern Recognition., vol. 2, pp: 891-896.

12. Sullivan, F. O., 1986. A statistical perspective on ill-posed inverse problems. Stat. Sci., 1(4), 502-518.

13. Swets, D. L., Weng, J. J., 1996. Using discriminant eigen features for image Retrieval. IEEE Trans. Pattern Anal. Mach. Intell., 8, 831-836.

14. Titterington, D. M., 1981. Common structure of smoothing techniques in statistics. Int. Stat. Rev., 141-170.

15. Yang, J., Yang, J., 2003. Why can lda be performed in pca transformed space. Pattern Recogn., 36(2), 563-566.

16. Yu, H., Yang, J., 2001. A direct LDA algorithm for high-dimensional data with application to face recognition. Pattern Recogn., 34(10), 2067-2070.

17. Zhao, W., Chellappa, R., Phillips, P. J., 1999. Subspace linear discriminant analysis for face recognition. Computer Vision Laboratory, Center for Automation Research, University of Maryland.

18. Zheng, W., Zhao, L., Zou, C., 2004. An efficient algorithm to solve the small sample size problem for LDA. Pattern Recogn., 37(5), 1077- 1079

19. Zheng, W., Zou, C., Zhao, L., 2004. Real-time face recognition using gram-schmidt orthogonalization for LDA. In Proceedings of the 17th IEEE International Conference on Pattern Recognition, 2, pp: 403- 406.