A Covariance Estimator for Small Sample Size Classification Problems and Its Application to Feature Extraction

Bor-Chen Kuo and David A. Landgrebe

March 8, 2012

Outline

- Introduction
- Previous methods for regularization
- Mixed Leave-one-out covariance (Mixed-LOOC) estimators
  - Mixed-LOOC 1
  - Mixed-LOOC 2
- Experiment design for comparing LOOC, Mixed-LOOC1 and Mixed-LOOC2
- Discriminate analysis feature extraction based on Mixed-LOOC / Experiments
- Conclusions

Motivation

- High-dimensional data (such as multi-spectral images) usually are classified using quadratic maximum-likelihood algorithm (ML)
- Classes must be modeled by a set of subclasses, each of which is described as a mean vector and a covariance matrix, whose parameters are learnt with ML
- When the number of training samples is low compared to the dimensionality, problems arise. Approaches
  - dimensionality reduction by feature extraction or feature selection
  - regularization of sample covariance matrix
  - structurization of a true covariance matrix described by a small number of parameters

Leave-one-out Covariance (LOOC)

$$\hat{\Sigma}_i(\alpha_i) = \begin{cases} (1-\alpha_i)\mathrm{diag}(S_i) + \alpha_i S_i & 0 \le \alpha_i \le 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S, & 1 \le \alpha_i \le 2 \\ (3-\alpha_i)S + (\alpha_i - 2)\mathrm{diag}(S), & 2 \le \alpha_i \le 3. \end{cases} \quad (1)$$

The mean of class i, without sample k, is

$$m_{i/k} = \frac{1}{N_i - 1} \sum_{\substack{j=1 \\ j \neq k}}^{N_i} x_{i,j}$$

The sample covariance of class i, without sample k, is

$$\Sigma_{i/k} = \frac{1}{N_i - 2} \sum_{\substack{j=1 \\ j \neq k}}^{N_i} (x_{i,j} - m_{i/k})(x_{i,j} - m_{i/k})^T \quad (2)$$

and the common covariance, without sample k from class i, is

$$S_{i/k} = \left(\frac{1}{L} \sum_{\substack{j=1 \\ j \neq i}}^{L} \Sigma_j \right) + \frac{1}{L} \Sigma_{i/k}. \quad (3)$$

The proposed estimate for class i, without sample k, can then be computed as follows:

$$C_{i/k}(\alpha_i) = \begin{cases} (1-\alpha_i)\mathrm{diag}(\Sigma_{i/k}) + \alpha_i \Sigma_{i/k}, & 0 \le \alpha_i \le 1 \\ (2-\alpha_i)\Sigma_{i/k} + (\alpha_i - 1)S_{i/k}, & 1 < \alpha_i \le 2 \\ (3-\alpha_i)S_{i/k} + (\alpha_i - 2)\mathrm{diag}(S_{i/k}), & 2 < \alpha_i \le 3. \end{cases} \quad (4)$$

The mixing parameter α_i is determined by maximizing the average leave-one-out log likelihood of each class

$$\mathrm{LOOL}_i = \frac{1}{N_i} \sum_{k=1}^{N_i} \ln[f(x_k | m_{i/k}, C_{i/k}(\alpha_i))]. \quad (5)$$

Bayesian LOOC (BLOOC)

$$\hat{\Sigma}_i(\alpha_i) = \begin{cases} (1-\alpha_i)\dfrac{\mathrm{tr}(S_i)}{p} I + \alpha_i S_i, & 0 \le \alpha_i \le 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S_p^*(t), & 1 \le \alpha_i < 2 \\ (3-\alpha_i)S + (\alpha_i - 2)\dfrac{\mathrm{tr}(S)}{p} I, & 2 < \alpha_i \le 3 \end{cases}$$

where t can be expressed as the function of α_i

$$t = \frac{(\alpha_i - 1)f_i - \alpha_i(p+1)}{2-\alpha_i},$$

where p is the dimensionality and $f_i = N_i - 1$, which represents the degree of freedom in Wishart distributions, and the pooled covariance matrices are determined under a Bayesian context and can be represented as

$$S_p^*(t) = \left[\sum_{i=1}^{L} \frac{f_i}{f_i + t - p - 1}\right]^{-1} \sum_{i=1}^{L} \frac{f_i S_i}{f_i + t - p - 1}. \quad (6)$$

Mixed-LOOC 1

$$\hat{\Sigma}_i(a_i, b_i, c_i, d_i, e_i, f_i) = a_i \frac{\mathrm{tr}(S_i)}{p} I + b_i \mathrm{diag}(S_i) + c_i S_i + d_i \frac{\mathrm{tr}(S)}{p} I + e_i \mathrm{diag}(S) + f_i S$$

where

$$a_i + b_i + c_i + d_i + e_i + f_i = 1 \quad \text{and} \quad i = 1, 2, \ldots, L \quad (7)$$

and

- L number of classes;
- p of dimensions;
- S_i covariance matrix of class i;
- S common covariance matrix (pooled).

The mixture parameters are determined by maximizing the average leave-one-out log likelihood of each class

$$\mathrm{LOOL}_i = \frac{1}{N_i} \sum_{k=1}^{N_i} \ln[f(x_k | m_{i/k}, \hat{\Sigma}_{i/k}(\theta_i))]$$

where

$$\theta_i = (a_i, b_i, c_i, d_i, e_i, f_i). \quad (8)$$

Mixed-LOOC 2

$$\hat{\Sigma}_i(\alpha_i) = \alpha_i A + (1 - \alpha_i) B \tag{9}$$

where $A = (\mathrm{tr}(S_i)/p)I$, $\mathrm{diag}(S_i)$, S_i, $(\mathrm{tr}(S)/p)I$, $\mathrm{diag}(S)$, or S, $B = S_i$, or $\mathrm{diag}(S)$ and α_i is close to 1. $B = S_i$, or $\mathrm{diag}(S)$ is chosen because if a class sample size is large, S_i will be a better choice. If total training sample size is less than the dimensionality, then the common (pooled) covariance S is singular but has much less estimation error than S_i. For reducing estimation error and avoiding singularity, $\mathrm{diag}(S)$ will be a good choice. The selection criteria is the log leave-one-out likelihood function

$$\mathrm{LOOL}_i = \frac{1}{N_i} \sum_{k=1}^{N_i} \ln[f(x_k | m_{i/k}, \hat{\Sigma}_{i/k}(\alpha_i))]. \tag{10}$$

Settings

- LOOC: $\alpha = [0, 0.25, 0.5, ..., 2.75, 3]$
- Mixed-LOOC1: values for the six parameters are in $[0, 0.25, 0.5, 0.75, 1]$
- Mixed-LOOC2: $\alpha = 0.05$
- Experiments 1 to 12 are based on simulated sets, randomly generated from two different mean vectors and covariances (same set in 1 to 6 and 7 to 12)
- Experiments 1 to 6 are balanced. Experiments 7 to 12 are unbalanced
- Dimensionality $p = 10, 30, 60$

Simulated databases

TABLE II
(a) ACCURACY OF SIMULATED DATA SETS ($p = 10$). (b) ACCURACY OF SIMULATED DATA SETS ($p = 30$).
(c) ACCURACY OF SIMULATED DATA SETS ($p = 60$). (d) ACCURACY OF REAL DATA SETS ($p = 191$)

Experiment	LOOC	Mixed-LOOC1	Mixed-LOOC2
1	0.8630 (0.0425)	0.8632 (0.0243)	0.8602 (0.0466)
2	0.7753 (0.0481)	0.8373 (0.0180)	0.8450 (0.0224)
3	0.8948 (0.0241)	0.8915 (0.0251)	0.8992 (0.0265)
4	0.8875 (0.0309)	0.8893 (0.0263)	0.8837 (0.0386)
5	0.9860 (0.0283)	0.9822 (0.0361)	0.9858 (0.0282)
6	0.9885 (0.0033)	0.9833 (0.0085)	0.9885 (0.0036)
7	0.8500 (0.0286)	0.8622 (0.0252)	0.8641 (0.0249)
8	0.8433 (0.0410)	0.8750 (0.0289)	0.8792 (0.0250)
9	0.9021 (0.0230)	0.9041 (0.0183)	0.9041 (0.0203)
10	0.8928 (0.0247)	0.8948 (0.0204)	0.8940 (0.0245)
11	0.9883 (0.0064)	0.9920 (0.0041)	0.9872 (0.0065)
12	0.9841 (0.0076)	0.9830 (0.0075)	0.9827 (0.0116)

(a)

Experiment	LOOC	Mixed-LOOC1	Mixed-LOOC2
1	0.8317 (0.0227)	0.8285 (0.0196)	0.8267 (0.0213)
2	0.7263 (0.0510)	0.8700 (0.0205)	0.8813 (0.0204)
3	0.8162 (0.0220)	0.8142 (0.0223)	0.8152 (0.0237)
4	0.7978 (0.0619)	0.7955 (0.0609)	0.7972 (0.0612)
5	0.9993 (0.0014)	0.9975 (0.0037)	0.9993 (0.0014)
6	0.9990 (0.0021)	0.9945 (0.0087)	0.9992 (0.0016)
7	0.8239 (0.0345)	0.8469 (0.0154)	0.8504 (0.0171)
8	0.8718 (0.0311)	0.9210 (0.0130)	0.9189 (0.0118)
9	0.8228 (0.0274)	0.8343 (0.0206)	0.8241 (0.0268)
10	0.8326 (0.0162)	0.8370 (0.0186)	0.8313 (0.0156)
11	0.9976 (0.0021)	0.9994 (0.0008)	0.9984 (0.0018)
12	0.9953 (0.0059)	0.9991 (0.0007)	0.9978 (0.0047)

(b)

Experiment	LOOC	Mixed-LOOC1	Mixed-LOOC2
1	0.7378 (0.0540)	0.7607 (0.0259)	0.7605 (0.0287)
2	0.6578 (0.0631)	0.8792 (0.0213)	0.8882 (0.0175)
3	0.7632 (0.0265)	0.7615 (0.0235)	0.7583 (0.0281)
4	0.7483 (0.0324)	0.7473 (0.0308)	0.7435 (0.0288)
5	1.0000 (0.0000)	0.9998 (0.0005)	1.0000 (0.0000)
6	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
7	0.7820 (0.0327)	0.8098 (0.0229)	0.8120 (0.0192)
8	0.8876 (0.0219)	0.9401 (0.0075)	0.9400 (0.0073)
9	0.7947 (0.0216)	0.8024 (0.0150)	0.7958 (0.0203)
10	0.7802 (0.0302)	0.7932 (0.0277)	0.7837 (0.0275)
11	0.9988 (0.0021)	0.9997 (0.0011)	0.9997 (0.0011)
12	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)

(c)

Real Data Set	LOOC	Mixed-LOOC2
Cuprite	0.7743 (0.1372)	0.9524 (0.0117)
Jasper Ridge	0.9864 (0.0042)	0.9849 (0.0019)
Indian Pine	0.7612 (0.0127)	0.7625 (0.0144)
DC Mall	0.7831 (0.0455)	0.7858 (0.0431)

(d)

Real database

TABLE III
THE MEAN ACCURACIES AND STANDARD DEVIATIONS OF EXPERIMENTS

Real Data Set	Exp17 DAFE+GC	Exp18 DAFE-Mix2+GC	Exp19 DAFE-Mix2+GC-Mix2
Cuprite	0.8943 (0.0205)	0.9474 (0.0194)	0.9627 (0.0196)
Jasper Ridge	0.9127 (0.0243)	0.9782 (0.0120)	0.9876 (0.0036)
Indian Pine	0.5727 (0.0156)	0.7547 (0.0316)	0.7562 (0.0191)
DC Mall	0.7392 (0.0530)	0.8691 (0.0282)	0.8600 (0.0345)

DAFE

The purpose of discriminate analysis feature extraction (DAFE) is to find a transformation matrix A such that the class separability of transformed data $Y = A^T X$ is maximized. Usually within-class, between-class, and mixture scatter matrices are used to formulate the criteria of class separability. A within-class scatter matrix is expressed by [9]

$$S_w = \sum_{i=1}^{L} P_i E\{(X - m_i)(X - m_i)^T | \omega_i\} = \sum_{i=1}^{L} P_i \Sigma_i \quad (11)$$

where L is the number of classes and P_i and m_i are the prior probability and mean vector of the class i, respectively.

A between-class scatter matrix is expressed as

$$\begin{aligned} S_b &= \sum_{i=1}^{L} P_i (m_i - m_0)(m_i - m_0)^T \\ &= \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} P_i P_j (m_i - m_j)(m_i - m_j)^T \end{aligned} \quad (12)$$

where m_0 represents the expected vector of the mixture distribution and is given by

$$m_0 = E\{X\} = \sum_{i=1}^{L} P_i m_i. \quad (13)$$

Let $Y = A^T X$, then we have

$$S_{wY} = A^T S_{wX} A \quad \text{and} \quad S_{bY} = A^T S_{bX} A. \quad (14)$$

The optimal features are determined by optimizing the criterion given by

$$J_1 = \text{tr}(S_{wY}^{-1} S_{bY}). \quad (15)$$

The optimum A must satisfy

$$(S_{wX}^{-1} S_{bX}) A = A (S_{wY}^{-1} S_{bY}). \quad (16)$$

This is a generalized eigenvalue problem [10] and usually can be solved by the QZ algorithm. But if the covariance is singular, the result will have a poor and unstable performance on classification. In this section, the ML covariance estimate will be replaced by Mixed-LOOC when it is singular. Then the problem will become a simple eigenvalue problem.

Results

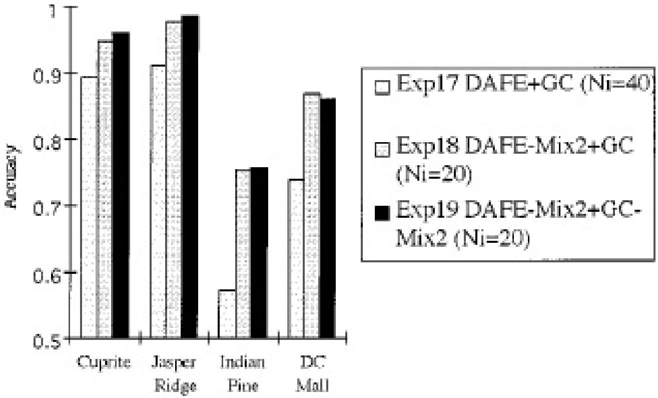

Fig. 4. The mean accuracies of three classification procedures.