Appl Intell (2012) 36:542–552
DOI 10.1007/s10489-010-0274-8

Multidimensional local spatial autocorrelation measure for integrating spatial and spectral information in hyperspectral image band selection

Zheng Du · Young-Seon Jeong · Myong K. Jeong · Seong G. Kong

Published online: 22 February 2011

Abstract Hyperspectral band selection aims at the determination of an optimal subset of spectral bands for dimensionality reduction without loss of discriminability. Many conventional band selection approaches depend on the concept of "statistical distance" measure between the probability distributions characterizing sample classes. However, the maximization of separability does not necessarily guarantee that a classification process results in the best classification accuracies. This paper presents a multidimensional local spatial autocorrelation (MLSA) measure that quantifies the spatial autocorrelation of the hyperspectral image data. Based on the proposed spatial measure, a collaborative band selection strategy is developed that combines both spectral separability measure and spatial homogeneity measure for hyperspectral band selection without losing the spectral details useful in classification processes. The selected band subset by the proposed method shows both larger separability between classes and stronger spatial similarity within class. Case studies in biomedical and remote sensing applications demonstrate that the MLSA-based band selection approach improves object classification accuracies in hyperspectral imaging compared with conventional approaches.

Keywords Hyperspectral band selection · Multidimensional local spatial autocorrelation · Spatial and spectral information

1 Introduction

Hyperspectral imaging sensors collect image intensity information in a number of narrow spectral bands that often range from the visible to near infrared spectra. A hyperspectral image can be characterized by a three-dimensional volume of data in spatial and spectral spaces [13]. The increased number of spectral bands provides more spectral details for better discriminating power in hyperspectral image analysis, but with higher dimensionality of data. Such a large amount of hyperspectral image data raises problems in storage and transmission, which make real-time computer processing of hyperspectral image data a challenging task. Since hyperspectral imaging sensors acquire band images in narrow, adjacent spectral bands, the resulting high-dimensional data sets are often highly correlated and contain redundant information. A reduced number of spectral bands may contain sufficient information to represent the entire dataset. The reduction of spectral bands results in both a decrease in computational time and an increase in the classification accuracy. The band selection procedure finds the small subset

Z. Du
Department of Electrical and Computer Engineering, The University of Tennessee, Knoxville, TN 37996-2100 USA
e-mail: zhengdu@gmail.com

Y.-S. Jeong
Department of Industrial and Systems Engineering, Rutgers University, Piscataway, NJ 08854-8003, USA
e-mail: ysjeong@eden.rutgers.edu

M.K. Jeong (✉)
Department of Industrial and Systems Engineering & RUTCOR, Rutgers University, Piscataway, NJ 08854-8003, USA
e-mail: mjeong@rci.rutgers.edu

M.K. Jeong
Department of Industrial and Systems Engineering, KAIST, Daejon 305-701, Korea

S.G. Kong
Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA 19122, USA
e-mail: skong@temple.edu

of bands that are relevant to the target process. Benefits of the band selection include reducing the number of bands, removing irrelevant, redundant, or noisy data, speeding up processing time, and improving classification performance [14, 16].

Many band selection approaches in the literature depend on the concept of "statistical distance" measure between the probability distributions characterizing the sample classes. Supervised band selection techniques involve criterion functions, which aim at evaluating the separability of classes for a given subset of spectral bands. Criterion functions are usually based on separability indices that express the goodness of each band subset. Bruzzone et al. [3] proposed a band selection method based on Jeffreys-Matusita (JM) distance. Canonical analysis (CA) maximizes the between-class scatter matrix while minimizing the within-class scatter matrix to achieve the maximum class discrimination through discriminating power defined in Tu et al. [22]. Using the eigenvalues and eigenvectors generated by the CA, a loading factor matrix can be calculated for each spectral band and be used as the classification capability of that band. The Bhattacharyya distance [7] and the Mahalanobis distance [23] have also been widely used for band selection in multispectral image analysis. These measures rank the bands to separate discriminative bands from irrelevant and redundant bands. Divergence [21] takes into account the correlation that exists among the various selected bands and influences on classification capability of selected spectral bands. Du et al. [5] proposed a band selection method based on recursive divergence to overcome the computational restrictions of a divergence approach. Band selection methods based on statistical distance measures maximize spectral separability. However, the maximization of separability does not guarantee a maximum classification accuracy.

A basic property of spatially located data is that the set of values are likely to be related over space [6, 9]. In hyperspectral imaging, there will be some degree of dependency between pixels. Hyperspectral image analysis techniques such as enhancement and classification can be treated as an attempt to make the spatial pattern clearer. In other words, pixels after process should on average be more similar to neighboring pixels than those pixels that are far away, a characteristic known as spatial autocorrelation. The spatial autocorrelation for the image provides an excellent measure of spatial information in the image. Identifying band combinations with the highest spatial autocorrelation should not only increase the accuracy of the spectral representation of the objects, but also increase their spatial representation and suppress visually distracting.

This paper proposes a new band selection method based on the integration of spatial and spectral information for hyperspectral image. The major impediment to the combined spatial and spectral information for hyperspectral image processing is that most spatial methods were developed for single image band. Based on the traditional single-image based local Geary measure, this paper develops a multidimensional local spatial autocorrelation (MLSA) measure that quantifies the spatial autocorrelation of hyperspectral images. Based on the proposed spatial measure, this paper develops a collaborative band selection strategy that combines both the spectral separability measure (divergence) and the spatial homogeneity measure for hyperspectral band selection task. Case studies in different applications demonstrate that the proposed method based on the MLSA measure improves object classification accuracies in hyperspectral imaging compared with combined spatial and spectral approaches.

2 Multidimensional local spatial autocorrelation

2.1 Spatial autocorrelation

Spatial autocorrelation is defined to measure the spatial dependence over a study area. Spatial autocorrelation can be measured using the Moran [4], the Geary [4], and local indicators spatial association (LISA) [1]. There exist two types of measures: global measures that provide a single value that summarize the level of spatial autocorrelation with respect to the whole region, and local measures that provide a value for each location with respect to its neighborhood.

2.1.1 Global spatial autocorrelation measures

The Moran I measures the global spatial autocorrelation

$$I = \frac{n}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \quad (1)$$

where x_i denotes the observed value at location i, $\bar{x}$ is the sample mean of the x_i over the n locations. w_{ij} is the (i, j)-th element of a spatial weight matrix. Here we consider symmetric binary weights, with ones if location j is contiguous to location i, and zeros otherwise. Moran's I statistic is approximately equal to zero when there is no spatial autocorrelation. $I > 0$ indicates a positive spatial autocorrelation, while $I < 0$ a negative one [4]. As an alternative approach to measuring spatial association, Geary's c statistic is defined as:

$$c = \frac{(n-1)}{2n\sum_{i=1}^{n}(x_i - \bar{x})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \quad (2)$$

The Geary statistic is always positive and asymptotically normal. The Geary value of 1 means there is no spatial autocorrelation. A low value (between 0 and 1) indicates a positive spatial autocorrelation while a high value (greater than 1) indicates a negative spatial autocorrelation [20].

2.1.2 Local spatial autocorrelation

Global spatial autocorrelation measures highlight the average spatial dependence over a study area. However, the global measure is useful only when spatial dependence is relatively uniform over the study area. If the underlying spatial process is not stationary, global measures may not be representative. In this case, since the global measures generate only an average measure of spatial dependency, it tends to vague any significant local variation of spatial nonstationarity in the study area. Therefore, a global estimate can be uninformative and misleading. Hence, it is often more appropriate to measure the spatial dependence at a smaller area. To overcome these limitations, local indicators of spatial association (LISA) has been developed [1] to identify the location and spatial dependence within the study area. In contrast to global spatial methods, the LISA focuses on local variations within patterns of spatial dependence. In calculating local spatial association measures of the image data, each pixel receives a value quantifying its spatial dependence to its neighbors, where the neighborhood is determined by the weights matrix.

Among several local measures, two popular measures are local Moran and local Geary statistics. The local Moran statistic for each observation i is defined as

$$I_i = \frac{(x_i - \bar{x}) \sum_{j=1}^{n} w_{ij}(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2/n} \tag{3}$$

for any $i = 1, \ldots, n$. The local Moran's I_i measures the joint covariance of neighboring localities. A local Geary statistic for each observation i is defined as follows

$$c_i = \frac{\sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2/n}. \tag{4}$$

Unlikely to local Moran's I_i, local Geary statistic is the weighted sum of the squared differences between location i and locations j. One property of the local Moran and Geary statistic is that they can be associated with the global statistics (Moran I and Geary c, respectively) and can be used to estimate the contribution of individual statistics to the corresponding global statistics.

2.2 Multidimensional local spatial autocorrelation (MLSA)

Most existing autocorrelation measures are developed for single band images. While the pixels in hyperspectral image are usually multi-dimensional vectors, the classical local Geary statistic cannot be directly applied to hyperspectral image data. In this section, a new measure, named multidimensional local spatial autocorrelation (MLSA) measure, is proposed to extend the single-image based local Geary measure to high dimensional data.

Let us now consider a hyperspectral image $\boldsymbol{H}$, defined on the K-dimensional space, where K indicates the number of spectral channels. We denote $\boldsymbol{x}_i$ the observed value at location i, and note that $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik}, \ldots, x_{iK})^t$ is represented as a column vector where k is the dimension of the data or the spectral band of the hyperspectral image. Suppose that $\boldsymbol{x}_i, \boldsymbol{x}_j \in N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, i.i.d, then MLSA can be defined as

$$c_i^V = \sum_j w_{ij}(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j) \tag{5}$$

where the superscript V in c_i^V indicates a vector, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and covariance matrix for $\boldsymbol{x}$, respectively. Some expectations of a random variable $\boldsymbol{x}_i$ are

$$E\left(\boldsymbol{x}_i \boldsymbol{x}_i^T\right) = \mathrm{Var}(\boldsymbol{x}_i) + \boldsymbol{\mu}\boldsymbol{\mu}^T = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T \tag{6}$$

$$\begin{aligned} &E\left[(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T\right] \\ &\quad = E\left[\boldsymbol{x}_i \boldsymbol{x}_i^T - \boldsymbol{x}_i \boldsymbol{x}_j^T - \boldsymbol{x}_j \boldsymbol{x}_i^T - \boldsymbol{x}_j \boldsymbol{x}_j^T\right] \\ &\quad = 2E\left[\boldsymbol{x}_i \boldsymbol{x}_i^T\right] - 2E(\boldsymbol{x}_i)E\left(\boldsymbol{x}_i^T\right) \\ &\quad = 2\,\mathrm{Var}(\boldsymbol{x}_i) + 2\boldsymbol{\mu}\boldsymbol{\mu}^T - 2E(\boldsymbol{x}_i)E\left(\boldsymbol{x}_i^T\right) \\ &\quad = 2\boldsymbol{\Sigma} \end{aligned} \tag{7}$$

The following lemma shows the moments of the proposed c_i^V under the null hypothesis of no spatial autocorrelation (see Appendix for its proof).

Lemma 1 *For a hyperspectral image $\boldsymbol{H}$, if $\boldsymbol{x}_i, \boldsymbol{x}_j \in N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and are i.i.d, then, under the null hypothesis of no spatial autocorrelation,*

(i) *Expected value for c_i^V is $2w_i K$,*
(ii) *Variance for c_i^V is $w_{i(2)}(4K^2 + 8K) + (w_i^2 - w_{i(2)}) \times (4K^2 + 2K) - (2w_i K)^2$.*

3 Collaborative band selection

Object classification based only on spectral information does not guarantee to obtain the most accurate results. The spatial information is a useful supplement to increase the classification accuracy [12]. In this section, we present a collaborative band selection (CBS) method that combines both spectral separability measure and spatial homogeneity measure of hyperspectral band selections. The CBS algorithm consists of three major steps of computation.

At the first step, recursive divergence proposed is used to measure the class separability of data samples for possible subset combinations [5]. Suppose that signal classes are characterized by p-dimensional multivariate normal distributions: $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean vector

and covariance matrix of class ω_i, respectively. Then, the divergence between these two classes is given by

$$D_{ij}(\boldsymbol{x}) = \frac{1}{2}\mathrm{tr}\big[\big(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1}\big)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\big] + \frac{1}{2}\mathrm{tr}\big[(\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j)\big(\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_i^{-1}\big)\big] \tag{8}$$

where tr denotes the trace operator of a matrix. In addition, for multi-class problems, the transformed divergence (TD) [8] instead of classical divergence should be adopted as the criterion. The TD gives an exponentially decreasing weight to increasing distances between the classes, therefore may result in a better performance. The TD is defined as

$$\mathrm{TD}_{ij}(\boldsymbol{x}) = 2\left[1 - \exp\left(-\frac{D_{ij}(\boldsymbol{x})}{8}\right)\right] \tag{9}$$

The basic idea of recursive divergence is to build up a set of d spectral bands incrementally, starting with the empty set. That is, the search algorithm constructs the set of spectral bands at the ith stage of the algorithm from that at the $(i-1)$th stage by the addition of a spectral band from the current optimal set. The divergence criterion (8) at stage i can be evaluated by updating its value already calculated for stage $(i-1)$ instead of computing the divergence. This results in substantial computational savings.

Let $D_{ij}(\boldsymbol{x}_p^*)$ be the divergence with p selected bands and $D_{ij}(\boldsymbol{x}_p^*, x_{p+1}^*)$ the divergence with the additional band x_{p+1}^*. Suppose the additional band x_{p+1}^* has mean μ_k^*, variance σ_k^2; and the covariance vector between x_{p+1}^* and the elements of $\boldsymbol{x}_p$, $\boldsymbol{z}_k$ for class k $(= i$ or $j)$. Then the new mean vectors and new covariance matrix are $\boldsymbol{\mu}_k^v = (\boldsymbol{\mu}_{k,p}^*; \mu_k^*)^T$, $(k = i$ or $j)$ and

$$\boldsymbol{\Sigma}_{k,p+1} = \begin{pmatrix} \boldsymbol{\Sigma}_{k,p} & \boldsymbol{z}_k \\ \boldsymbol{z}_k^t & \sigma_k^2 \end{pmatrix} \tag{10}$$

The divergence with the additional of a band x_{p+1}^* can be recursively calculated in an efficient way as follows:

$$D_{ij}(\boldsymbol{x}_p^*, x_{p+1}^*) = D_{ij}(\boldsymbol{x}_p^*) + \Delta_{ij}(x_{p+1}^*) \tag{11}$$

where $\Delta_{ij}(x_{p+1}^*)$ is the incremental divergence due to the addition of a band x_{p+1}^*, and can be calculated by the following formulae:

$$\begin{aligned}\Delta_{ij}(x_{p+1}^*) = {} & \frac{1}{2\delta_i}\big[(\mu_i^* - \mu_j^*) - (\boldsymbol{\mu}_{i,p}^* - \boldsymbol{\mu}_{j,p}^*)^T\gamma_i\big]^2 \\ & + \frac{1}{2\delta_j}\big[(\mu_i^* - \varpi_j^*) - (\boldsymbol{\mu}_{i,p}^* - \boldsymbol{\mu}_{j,p}^*)^t\gamma_j\big]^2 \\ & + \frac{1}{2}\mathrm{tr}\big[(\boldsymbol{\Sigma}_{i,p+1} - \boldsymbol{\Sigma}_{j,p+1}) \\ & \times (\delta_i^{-1}\gamma_i\gamma_i^t + \delta_j^{-1}\gamma_j\gamma_j^t)\big] \\ & + (\boldsymbol{z}_i^t - \boldsymbol{z}_j^t)(\delta_i^{-1}\gamma_i - \delta_j^{-1}\gamma_j) \\ & + (\sigma_i^2 - \sigma_j^2)(\delta_i - \delta_j)\end{aligned} \tag{12}$$

where $\gamma_k = \boldsymbol{\Sigma}_{k,p}^{-1}\boldsymbol{z}_k$ and $\delta_k = \sigma_k^2 - \boldsymbol{z}_k^t\boldsymbol{\Sigma}_{k,p}^{-1}\boldsymbol{z}_k$ [5].

The divergence values are sorted and ranked in a descending order for all combination, and then the CBS algorithm selects several subsets which have the largest divergence value. Since the classifier makes decision based on the spectral similarity, hence the large divergence value usually indicates higher classification accuracy. The algorithm, which chooses several band combinations with the largest divergence values, can improve the classification performance.

At the second step, the algorithm attempts to improve the classification accuracy by integrating spatial information. At this step, the multidimensional local spatial autocorrelation measure for those selected subsets in the previous step is calculated. By adopting the proposed MLSA measure, it is possible to efficiently calculate the spatial autocorrelation for hyperspectral images. From the training samples, the average MLSA measure for class ω_l of specific band subset (say $\boldsymbol{U}$, $\boldsymbol{U} = [\lambda_1, \ldots, \lambda_p]$) is calculated as:

$$C_l(\boldsymbol{x}_{\boldsymbol{U}}) = \frac{1}{N_l}\sum_{i=1}^{N_l} c_i^V(\boldsymbol{x}_{\boldsymbol{U}}) \tag{13}$$

The average MLSA measure for all training samples is

$$C(\boldsymbol{x}_{\boldsymbol{U}}) = \sum_{l=1}^{L} C_l(\boldsymbol{x}_{\boldsymbol{U}}) \tag{14}$$

The third step is to combine the spectral information (divergence value) with the spatial information (average MLSA value). The ratio between divergence and average MLSA value is used to combine the divergence and spatial autocorrelation in this step:

$$DC(\boldsymbol{x}_{\boldsymbol{U}}) = D(\boldsymbol{x}_{\boldsymbol{U}})/C(\boldsymbol{x}_{\boldsymbol{U}}) \tag{15}$$

where $\boldsymbol{U}$ is a specific band subset. For the divergence measure, the larger value indicates more separation between two classes. While for the MLSA measure, the smaller value means stronger spatial similarity. For (15), the band subset, which has large class separability and strong spatial similarity, will yield a bigger output value. The rule to find the optimal subset at this step is defined as:

$$\boldsymbol{U}^* = \arg\max_{\boldsymbol{U}} DC(\boldsymbol{x}_{\boldsymbol{U}}) = \arg\max_{\boldsymbol{U}}\big[D(\boldsymbol{x}_{\boldsymbol{U}})/C(\boldsymbol{x}_{\boldsymbol{U}})\big] \tag{16}$$

Remark 1 The CBS method combines both the divergence measure and the local spatial autocorrelation measure, thus the complexity of CBS method is the sum of the complexities of two methods. The computational complexity of two-class divergence is $O(K)$, where K is the total number of

bands [2]. For multi-class divergence, the complexity time is $O(K^*C(C-1)/2)$, where C is the total number of classes. On the other hands, the algorithm complexity for the local spatial autocorrelation measure is $O(N)$, where N is the number of pixels. The total complexity of CBS method is $O(K^*C(C-1)/2)+O(N)$.

The procedure for the CBS method is described as follows.

The Collaborative Band Selection Algorithm

Input: *a set of spectral bands* $\boldsymbol{\Lambda}=[\lambda_1,\lambda_2,\ldots,\lambda_K]$, *training sample set* $\boldsymbol{X}$ *User defined value* p
Output: *selected optimal band subset* $\boldsymbol{O}$

1. *Set* $\boldsymbol{O}$ *to the empty set*.
2. *Exhaustively calculate divergence* $D(\lambda_k)$ *by* (8) *for all bands in* $\boldsymbol{\Lambda}$. *Sort and rand the* $D(\lambda_k)$ *in a descendent order. Find the largest p divergence values and corresponding subsets*.
3. *Calculate the average multidimensional local spatial autocorrelation measure for these p subsets*.
4. *Calculate the combination value according to* (15).
5. *Select the band having the largest combination value. Add it to the selected band set* $\boldsymbol{O}$ *and remove it from* $\boldsymbol{\Lambda}$.
6. *If stopping criterion is met, then stop and output the selected band set* $\boldsymbol{O}$. *Otherwise go to Step* 2.

4 Experimental results for band selection

A series of experiments were carried out to compare the performance of the proposed CBS method with existing band selection methods. The maximum likelihood classifier (MLC) is used to get the pixel classification results. Two real-life case studies using hyperspectral imagery are investigated, namely detection and identification of tumor on poultry carcasses [5, 10], and crops in an agricultural filed.

4.1 Hyperspectral imaging for food safety

Hyperspectral images obtained from the Instrumentation and Sensing Laboratory (ISL) consist of 460 × 400 pixels with 65 spectral bands. The spectral band has a discrete value from the wavelength λ_1 (442.9 nm) to λ_{61} (710.7 nm). The sample poultry carcasses were placed on a tray painted with a non-fluorescent flat black paint to minimize background scattering in a darkened room. The speed of the conveyer belt was adjusted based on the predetermined CCD exposure time and data transfer rate. The tumors on the poultry carcass are verified and labeled by a Food Safety and Inspection Service (FSIS) veterinarian (shown in Fig. 1). From the training data cube, we extract 1500 pixels from different tumor areas and 5000 pixels from different normal tissue areas. Twenty percent of these samples (300 tumor pixels and 1000 normal pixels) are used in training, and the other are used as the test dataset.

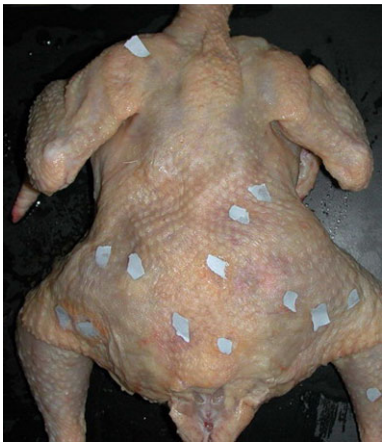
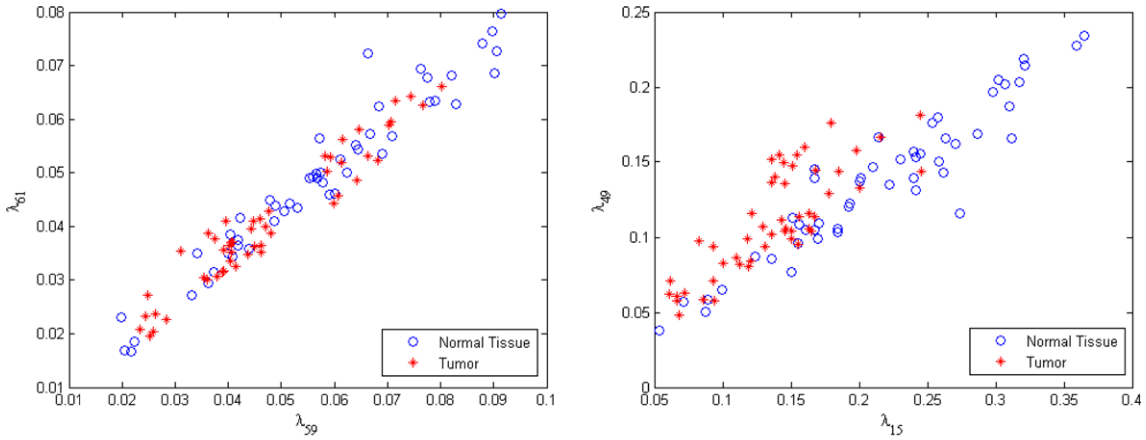


Fig. 1 Labeled tumors

Figure 2 shows the band combinations with minimum and maximum divergence values for two bands case. This result is obtained from exhaustively calculating all combinations with two bands. Bands λ_{59} and λ_{61} gives minimum divergence value 0.8262 among all 2080 combinations. As shown in Fig. 2(a), the normal tissue and tumor pixels are highly overlapped. That means if these two bands are used for classification, it will generate plenty of misclassified pixels. While for bands λ_{15} and λ_{49}, which gives the largest divergence value 10.8444, the tumor and normal tissue are more separate than the previous one. Figure 2 illustrates the relationship between divergence value and class separability. The larger the divergence value indicates more separability between two classes.

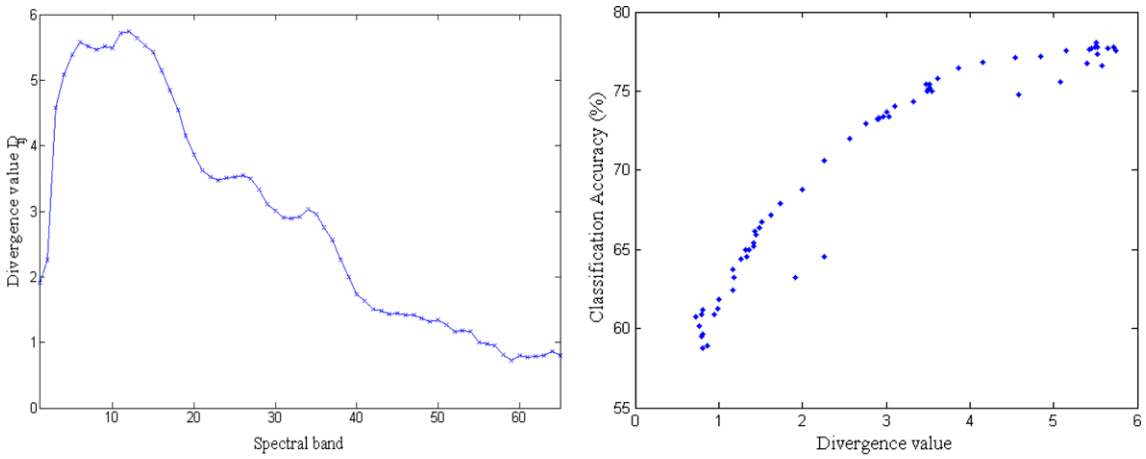
In case of band selection for 1-band, we choose the best one band from total 65 spectral bands. Figure 3(a) shows the divergence value for all 65 spectral bands. The divergence has a large peak value at approximately band λ_{12} with divergence value of 5.7454. If we use only the divergence value as the criterion, band λ_{12} will be chosen since it has the largest divergence value among all 65 bands. Figure 3(b) plots the classification accuracy corresponding with different divergence values. The maximum value of accuracy happens at band λ_9, which has divergence value 5.5088. These plots suggest that, in general, the larger divergence value usually result higher classification accuracy. However, maximizing the divergence does not guarantee to achieve the highest classification accuracy. We need to consider the spatial information as a useful supplement to increase the classification accuracy.

Table 1 shows the divergence values, average local Geary values and accuracy for ten bands. These ten bands have the largest divergence values among all 65 bands. The av-



(a) Minimum divergence value: 0.8262 (b) Maximum divergence value: 10.8444

Fig. 2 Minimum and maximum divergence values for 2-band case



(a) Divergence value for 1-band case (b) Divergence value vs. classification accuracy for testing data

Fig. 3 Divergence value and accuracy for 1-band case

Table 1 Divergence (D), average local Geary (C) and accuracy (Acc) for 1-band case

Band	12	11	13	6	14	7	9	10	8	15
D	5.7454	5.7120	5.6459	5.5760	5.5281	5.5165	5.5088	5.4895	5.4581	5.4327
C	0.1079	0.1130	0.1022	0.1539	0.1000	0.1411	0.1032	0.1180	0.1318	0.0965
Acc	0.7755	0.7775	0.7767	0.7663	0.7759	0.7734	0.7803	0.7778	0.7773	0.7778

erage local spatial autocorrelation values are calculated with a 3×3 window. The RD method chooses λ_{12} as the optimal band for one-band case, and λ_{12} yields 77.55% classification accuracy. For the CBS method, the λ_{15} is chosen as the optimal band, which has 77.78%. The highest accuracy 78.03% is obtained for λ_9, which is only 0.25% higher than that obtained by CBS. Hence for one band case, the CBS method can obtain better performance than RD method.

In case of the 2-band case, two spectral bands will be selected from 65 bands. There are $65 \times 64/2 = 2080$ combinations in total. Figure 4 plots the classification accuracy corresponding with different divergence values. Table 2 shows the classification results for different combinations of 2-band subsets. The subset $[\lambda_{14}, \lambda_{65}]$ gives the highest accuracy rate of 83.31% obtained by exhaustive search. The subset $[\lambda_{15}, \lambda_{49}]$ has the largest divergence value among 2080 pos-

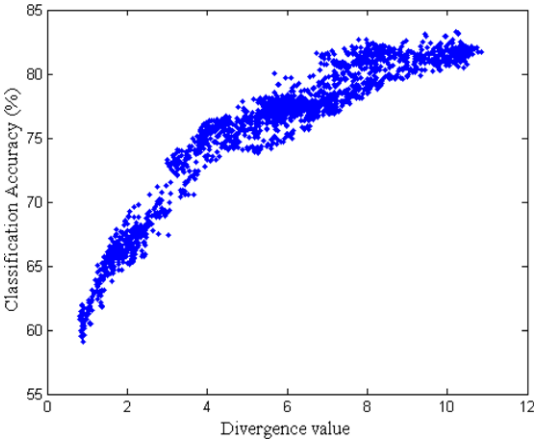


Fig. 4 Divergence value vs. classification accuracy for 2-band

Table 2 Divergence (D), average local Geary (C) and accuracy (Acc) for 2-band case

Band		D	C	Acc
14	65	10.2135	0.3110	0.8331
15	49	10.8444	0.3338	0.8169
12	46	9.8367	0.3212	0.8090
15	47	10.5840	0.3214	0.8196

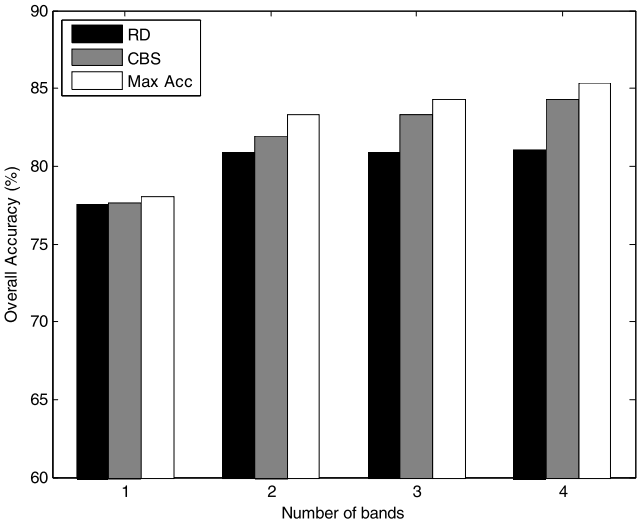
Table 3 Divergence (D), average local Geary (C) and accuracy (Acc) for 3-band case

Band			D	C	Acc
14	58	65	11.4371	0.7467	0.8427
12	46	2	10.3414	0.7331	0.8120
15	47	26	11.3698	0.7434	0.8329

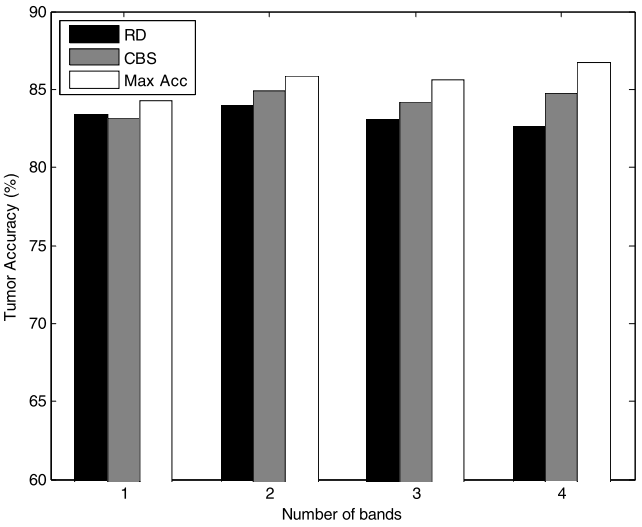
sible combinations. This subset yields accuracy of 81.69%, which is not the highest accuracy rate. The RD method chooses subset $[\lambda_{12}, \lambda_{46}]$, which has 80.90% accuracy. The CBS method selects subset $[\lambda_{15}, \lambda_{47}]$, and this subset can produce 81.96% accuracy. The RD and CBS method adopt the sequential forward search strategy, which keeps the selected bands from the previous stage, and adds a new band to make the subset optimal criterion. The CBS method could achieve the improved classification accuracy by considering the spatial information.

Table 3 shows the classification results for 3-band case. The highest accuracy 84.27% is yielded by subset $[\lambda_{14}, \lambda_{58}, \lambda_{65}]$ among all combinations of three bands, which is $65 \times 64 \times 63/6 = 43\,680$ combinations. The RD method chooses subset $[\lambda_{12}, \lambda_{46}, \lambda_{2}]$, which has 81.20% accuracy. The CBS method selects subset $[\lambda_{15}, \lambda_{47}, \lambda_{26}]$, and this subset produces 83.29% accuracy. The accuracy yielded by CBS method is very close to the optimal one.

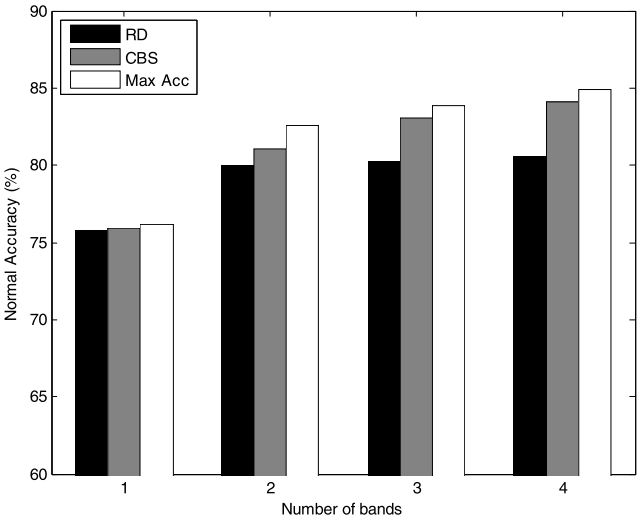
Figure 5 summarizes the classification accuracy for testing data set with different number of the selected bands from



(a) Overall accuracy



(b) Accuracy for Tumor



(c) Accuracy for Normal Tissue

Fig. 5 Classification accuracy with selected bands for chicken data

each procedure. The overall accuracy for selected bands is shown in Fig. 5(a). The CBS method produces higher classification accuracy than the RD method. For the tumor accuracy (shown in Fig. 5(b)), the RD and CBS method all generate high accuracies. This means the bands selected by RD and CBS method are all provided good discriminant ability for tumor pixels. But for normal tissue accuracy (shown in Fig. 5(c)), the CBS method produces the higher accuracy than RD. This suggests that the band selected by the CBS method will generate less false positive error than RD.

4.2 Indiana pine data

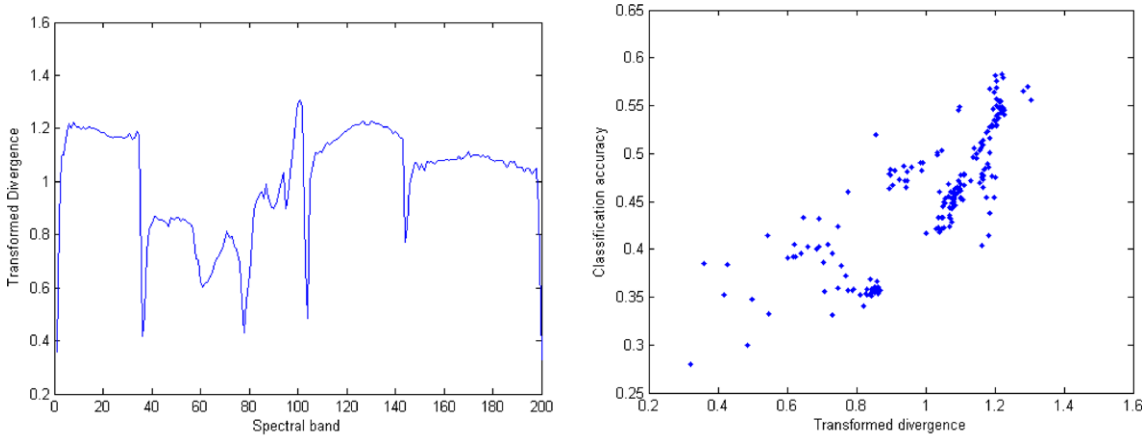
The previous hyperspectral dataset contains only two classes. In this experiment, we will test the proposed band selection method for multiple-class situations. The Indiana pine data used in this experiment contains four classes. The Indiana pine data is a well-known publicly available hyperspectral data set, which is used to investigate land use and can be downloaded from ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/. Data are delivered by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS), which measured pixel response in 224 bands in the 0.4 to 2.45 μm region of the spectrum with about 10 nm intervals, at a spatial resolution of 20 m, and covers an agricultural portion of North West Indiana. The dataset consists of 145 × 145 pixels in 220 contiguous spectral bands. Four of the 224 AVIRIS bands do not contain data, leaving 220 bands. The advantage of using this dataset is the availability of the reference image produced from field surveys, which may be used for accuracy assessment purposes. Similar to the earlier work on this dataset [15], twenty bands, $\lambda_{104} - \lambda_{108}$ (1.36–1.40 μm), $\lambda_{150} - \lambda_{163}$ (1.82–1.93 μm), and λ_{220} (2.50 μm), where the atmosphere is opaque have been omitted from the data set. In this paper, we use only a part of the 145 × 145 scene, called the subset scene for a size of 68 × 86. The subset scene contains four classes: Corn-notill, Soybean-notill, Soybean-mintill, and Grass-Trees, and over 75% of this scene are labeled. From the subset scene, 20% of the pixels were randomly chosen from the known ground truth of the four classes: Corn-notill, Soybean-notill, Soybean-min, and Grass-Trees. The remaining 80% of the known ground pixels in the scene are used as testing datasets. Table 4 lists the pixel numbers used as training and testing of each class.

Since this is a multi-class problem, we adopted transformed divergence (TD) instead of classical divergence as the criterion. Figure 6 shows the TD for one band case. Figure 6(a) plots the TD value for each band. The peak value appears at band λ_{101}. Figure 6(b) plots the TD value vs. pixel classification accuracy on testing data and shows that maximizing the divergence value does not guarantee the highest classification accuracy.

Experiments are carried out to compare the performance of our proposed band selection method with that of four band selection algorithms such as multiclass Bhattacharyya distance (BD) [19], multiclass Jeffries-Matusita (JM) distance [18], Relief [11], and Canonical analysis (CA) [22]. Figure 7 presents the comparison results of classification accuracy of six band selection methods. For all methods, the total accuracy improves as more spectral bands are added into the band subsets. The proposed CBS method achieves the best accuracy among others, indicating that spatial homogeneity measure is a positive supplement to increase the

Table 4 Indiana pine data used for training and testing

	Corn-notill	Grass/Tree	Soybean-notill	Soybean-min
# of training	202	146	145	385
# of testing	806	586	582	1541



(a) Transformed divergence value for each band

(b)TD vs. testing accuracy

Fig. 6 Transformed divergence for 1 band case

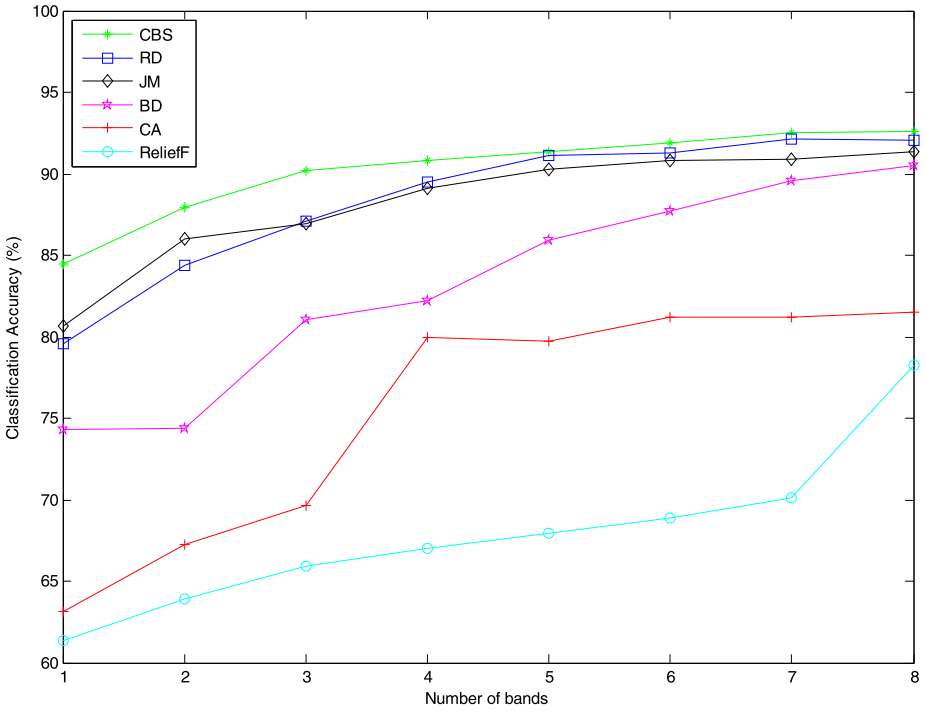


Fig. 7 Comparison of band selection method performance

classification accuracy. RD and JM show a similar performance, but much better than that of other two methods.

5 Conclusion

This paper has presented a new band selection method that integrates both the spectral and spatial information for hyperspectral image classification. While most band selection approaches only use the spectral information, spectral separability maximization does not guarantee a classification process that will produce the best visual results. This paper proposed a multidimensional local spatial autocorrelation measure for hyperspectral image data to measure the spatial autocorrelation. Based on the proposed spatial measure, this paper developed a collaborative band selection strategy that combines both the spectral separability measure and spatial homogeneity measure for hyperspectral band selection tasks. As shown by different applications, the proposed method effectively reduces the redundant bands with a minor classification accuracy loss.

In order to combine the spatial information with spectral information, first we need to have a criterion to measure the spatial information in the hyperspectral image. Although there exist many spatial statistic measures, most of them have only been developed for the single image band. In this paper, we proposed a MLSA measure to assess the spatial information for hyperspectral data. Additionally, in this paper, the sequential forward search strategy is used in band selection procedure. The advantage of this search strategy is computational efficients. But this strategy can only find a near-optimal solution.

Appendix A: Proof for Lemma 1

B.1 Proof of result (i)

The expected value for c_i^V can be denoted as

$$E(c_i^V) = E\left[\sum_j w_{ij}(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)\right] \tag{17}$$

By using the property of expectation, the (17) can be represented as

$$\begin{aligned} E(c_i^V) &= E\left[\sum_j w_{ij}(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)\right] \\ &= \sum_j w_{ij} E\left[(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)\right] \\ &= w_i E\left[(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)\right] \\ &= w_i E\left\{\mathrm{tr}\left[(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)\right]\right\} \\ &= w_i E\left\{\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T\right]\right\} \\ &= w_i \,\mathrm{tr}\left\{E\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T\right]\right\} \\ &= w_i \,\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1} E\left[(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T\right]\right\} \\ &= w_i \,\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(2\boldsymbol{\Sigma})\right\} \\ &= 2w_i K \end{aligned} \tag{18}$$

where K is the dimension of the features, $w_i = \sum_j w_{ij}$.

B.2 Proof of result (ii)

The variance for c_i^V can be denoted as

$$\mathrm{var}(c_i^V) = E\big[(c_i^V)^2\big] - E^2(c_i^V) \tag{19}$$

while

$$\begin{aligned} E\big[(c_i^V)^2\big] &= E\Bigg[\sum_{j\neq i} w_{ij}^2 (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j) \\ &\quad \times (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j) \\ &\quad + \sum_{k\neq l\neq i} \sum w_{ik}^2 (\boldsymbol{x}_i - \boldsymbol{x}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_k) \\ &\quad \times (\boldsymbol{x}_i - \boldsymbol{x}_l)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_l)\Bigg] \\ &= w_{i(2)} E\big[(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \\ &\quad \times \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)\big] \\ &\quad + (w_i^2 - w_{i(2)}) E\big[(\boldsymbol{x}_i - \boldsymbol{x}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_k) \\ &\quad \times (\boldsymbol{x}_i - \boldsymbol{x}_l)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_l)\big] \end{aligned} \tag{20}$$

where $w_{i(2)} = \sum_{j\neq i} w_{ij}^2$.

The first term in right hand side of (20) can be represented as [21]

$$\begin{aligned} &E\big[(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)\big] \\ &\quad = \mathrm{tr}(\boldsymbol{\Sigma}^{-1} 2\boldsymbol{\Sigma})\,\mathrm{tr}(\boldsymbol{\Sigma}^{-1} 2\boldsymbol{\Sigma}) + 2\,\mathrm{tr}(\boldsymbol{\Sigma}^{-1} 2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} 2\boldsymbol{\Sigma}) \\ &\quad = (2K)(2K) + 2(4K) = 4K^2 + 8K \end{aligned} \tag{21}$$

For the second term in (20),

$$\begin{aligned} &E\big[(\boldsymbol{x}_i - \boldsymbol{x}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_k)(\boldsymbol{x}_i - \boldsymbol{x}_l)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_l)\big] \\ &\quad = \mathrm{cov}\big((\boldsymbol{x}_i - \boldsymbol{x}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_k), (\boldsymbol{x}_i - \boldsymbol{x}_l)^T \\ &\qquad \times \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_l)\big) + E\big[(\boldsymbol{x}_i - \boldsymbol{x}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_k)\big] \\ &\qquad \times E\big[(\boldsymbol{x}_i - \boldsymbol{x}_l)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_l)\big] \end{aligned} \tag{22}$$

Using the following results [17],

$$\begin{aligned} &\mathrm{cov}\big((\boldsymbol{x}_i - \boldsymbol{x}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_k), (\boldsymbol{x}_i - \boldsymbol{x}_l)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_l)\big) \\ &\quad = 2K \end{aligned} \tag{23}$$

and

$$\begin{aligned} &E\big[(\boldsymbol{x}_i - \boldsymbol{x}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_k)\big] E\big[(\boldsymbol{x}_i - \boldsymbol{x}_l)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_l)\big] \\ &\quad = \mathrm{tr}(\boldsymbol{\Sigma}^{-1} 2\boldsymbol{\Sigma})\,\mathrm{tr}(\boldsymbol{\Sigma}^{-1} 2\boldsymbol{\Sigma}) = 4K^2 \end{aligned} \tag{24}$$

We can obtain

$$\begin{aligned} &E\big[(\boldsymbol{x}_i - \boldsymbol{x}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_k)(\boldsymbol{x}_i - \boldsymbol{x}_l)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_l)\big] \\ &\quad = 2K + 4K^2 \end{aligned} \tag{25}$$

The variance value turns out to be

$$\begin{aligned} \mathrm{var}(c_i^V) &= E\big[(c_i^V)^2\big] - E^2(c_i^V) \\ &= w_{i(2)}(4K^2 + 8K) + (w_i^2 - w_{i(2)})(4K^2 + 2K) \\ &\quad - (2w_i K)^2 \end{aligned} \tag{26}$$

References

1. Anselin L (1995) Local indicators of spatial association—LISA. Geogr Anal 27:93–115
2. Beigi M, Chang S, Ebadollahi S, Verma D (2009) Multi-scale temporal segmentation and outlier detection in sensor networks. IEEE International Conference on Multimedia and Expo, pp. 306–309
3. Bruzzone L, Roli F, Serpico S (1995) An extension of the Jeffreys-Matusita distance to multiclass cases for feature selection. IEEE Trans Geosci Remote Sens 33:1318–1321
4. Cliff A, Ord J (1981) Spatial Processes, Models and Applications. Pion, London
5. Du Z, Jeong M, Kong S (2007) Band selection of hyperspectral images for automatic detection of poultry skin tumors. IEEE Trans. Autom. Sci. Eng. 4(3):332–339
6. Getis A, Ord J (1992) The analysis of spatial association by use of distance statistics. Geogr Anal 24:189–206
7. Huang R, He M (2005) Band selection based on feature weighting for classification of hyperspectral data. IEEE Geosci Remote Sens Lett 2:156–159
8. Jensen J (1996) Introductory digital image processing: a remote sensing perspective. Prentice-Hall, Englewood Cliffs
9. Jeong Y, Kim S, Jeong M (2008) Automatic identification of defect patterns in semiconductor wafer maps using spatial correlogram and dynamic time warping. IEEE Trans Semicond Manuf 21(4):625–637
10. Kong S, Chen Y, Kim I, Kim M (2004) Analysis of hyperspectral fluorescence images for poultry skin tumor inspection. Appl Opt 43(4):824–833
11. Kononenko I (1994) Estimating attributes: analysis and extensions of RELIEF. In: Proceedings of seventh european conference on machine learning, pp 171–182
12. Kopisch M (1995) Spatial relations in technical domains. Appl Intell 5(4):351–366
13. Landgrebe D (2002) Hyperspectral image data analysis as a high dimensional signal processing problem. IEEE Signal Process Mag 19(1):17–28
14. Liu D, Setiono R (1998) Incremental feature selection. Appl Intell 9(3):217–230
15. Milenova B, Campos M (2005) Mining high-dimensional data for information fusion: a database-centric approach. In: Proceedings of 8th international conference on information fusion, vol 1, pp 7–14
16. Park J, Jeong M (2010) Recursive support vector censored regression for monitoring product quality based on degradation profiles. in press. Applied Intelligence
17. Schott J (2005) Matrix analysis for statistics, 2nd edn. Wiley, New Jersey

18. Serpico S, Moser G (2007) Extraction of spectral channels from hyperspectral images for classification purposes. IEEE Trans Geosci Remote Sens 45(2):484–495
19. Simin C, Rongqun Z, Wenling C, Hui Y (2009) Band selection of hyperspectral image based on Bhattacharyya distance. WSEAS Trans Inf Sci Appl 6(7):1165–1175
20. Stein A, van der Meer F, Gorte B (1999) Spatial statistics for remote sensing. Kluwer Academic, Dordrecht
21. Swain P, King R (1973) Two effective feature selection criteria for multispectral remote sensing. In: Proceedings of the first international joint conference on pattern recognition, pp 536–540
22. Tu T, Chen C, Wu J, Chang C (1998) A fast two-stage classification method for high-dimensional remote sensing data. IEEE Trans Geosci Remote Sens 36:182–191
23. Withagen P, Breejen F, Franken F, De Jong A, Winkel H (2001) Band selection from a hyperspectral data-cube for a real-time multispectral 3CCD camera. In: Proceedings of SPIE conference on algorithms for multispectral, hyperspectral, and ultraspectral imagery VII, vol 4381, pp 84–93