Forming Different-Complexity Covariance-Model Subspaces through Piecewise-Constant Spectra for Hyperspectral Image Classification

Are Charles Jensen[1] and Marco Loog[2]

[1] Department of Informatics, University of Oslo, Norway
[2] Pattern Recognition Laboratory, Delft University of Technology, The Netherlands

Abstract. A key factor in classifiers based on the normal (or Gaussian) distribution is the modeling of covariance matrices. When the number of available training pixels is limited, as often is the case in hyperspectral image classification, it is necessary to limit the complexity of these covariance models. An alternative to reducing the complexity uniformly over the whole feature space, is to form orthogonal subspaces and reduce the model complexity within them separately, e.g., forming full-complexity within-class, or interior-class, subspace models, and reduced-complexity exterior-class subspace models. We propose to use subspaces created by forming fewer and wider spectral bands, instead of the more general principal component analysis transform (PCA), in an attempt to exploit a-priori knowledge of the data to create more generalizable subspaces. We investigate the resulting classifiers by studying their performances on four hyperspectral data sets. On each data set, experiments where run using different training set sizes. The results indicate that the classifiers seem to benefit from using this more data-specific approach to forming subspaces.

1 Introduction

A recurring challenge in supervised classification of hyperspectral image data is that of handling the high number of per-object, or per-pixel, measurements (i.e., spectral bands) in combination with the often low number of samples available for training the classifiers. Typically we have about 100 to 200 spectral measurements per pixel, making the space in which we operate quite large and, usually, very sparsely sampled. As a result, when trying to build a statistical classifier, we most often have to resort to extremely simple models to avoid *overfitting* the training data. Even simple probability density functions (pdfs) like the normal distribution quickly turn out to be too complex, and, generally, we have to turn to dimensionality reduction, further restrictions of our model or other types of regularization, all preferably guided by some appropriate a-priori knowledge of the specific problem or data itself. The fewer samples we have available to train the classifiers, the more we rely on creating suitable restraints on our models to be able to harvest the spectral richness.

A. Heyden and F. Kahl (Eds.): SCIA 2011, LNCS 6688, pp. 186–195, 2011.

Many classifiers are based on modeling normal distributions, although differing greatly in how they impose a-priori stricture. In this paper we will focus on the approach of separating the feature space, principally differently for each class, into a *primary* and *secondary* orthogonal subspace, in which the complexity of how we model the two spaces differs. The primary space is meant to model a class' interior, or within, variance, while the secondary space models the exterior-class, or between-class, variance. The idea of such a separation for spectral data dates back to the work by Wold et al.[8] and Frank [2] in their SIMCA and DASCO approaches. If we choose the secondary space to cover a large part of the full feature space and let it be more simply modeled, we reduce the overall model complexity and hence limit the chance of overfitting.

In SIMCA, DASCO and common derivatives [7], the division of the feature space is based on an eigenvalue-decomposition (PCA) of the covariance matrix. A certain number of the eigenvectors corresponding to the highest eigenvalues are used to form the primary space, while the rest of the space is assumed to have equal variance in all directions, i.e., the remaining eigenvalues are set to their average value. By an eigenvalue-decomposition of the covariance matrix to form the primary space, we get the linear subspace containing the highest fraction of the total variance of the data, or, put another way, the subspace that can best represent the data in a squared error sense. However, there are no constraints or links to the data-generating process when the linear subspace is formed. This, in turn, could lead to an overfitting of the training samples, in the sense that it could make the primary space fail to represent the more generalizable within-class variance and give an artificially low variance in the secondary space.

In this paper, we suggest to replace the very general eigenvalue-decomposition with a more application-specific, or data-specific, approach to forming the primary and secondary subspaces. In particular, we propose to form primary spaces by finding the low-dimensional linear subspaces that, like the PCA, can best represent the data in a squared error sense, but with the restrictions that each basis vector corresponds to a single, wider spectral band, and that they together cover the whole spectrum. By enforcing this restriction, based on a-priori knowledge of the data, i.e., that they stem from samples of *continuous* curves, we should be able to obtain within-class subspaces that are less prone to overfitting.

In section 2 we recount the general normal distribution-based classifier, specify the feature-space separation formulation and give details of our proposed choices of how to define these subspaces. Details about the experiments, their results and a discussion on the findings can be found in section 3. Finally, section 4 gives some concluding remarks.

2 Model Formulation

Before we look at the specific models that we will investigate, we start by recapitulating the general formulation of the normal distribution-based classifiers.

2.1 Discriminant Analysis

Let $\mathbf{y}$ be a column vector containing the spectral band values of a single pixel. Now, by our assumption that each class follows a normal distribution, we end up with the following discriminant functions if we want to minimize the Bayes error [1]:

$$g_c(\mathbf{y}) = -\frac{1}{2}\log|\Sigma_c| - \frac{1}{2}(\mathbf{y}-\mu_c)'\Sigma_c^{-1}(\mathbf{y}-\mu_c) + \log\pi_c, \tag{1}$$

where μ_c, Σ_c^{-1} and π_c are the mean vector, the inverse covariance matrix (also called the precision matrix) and a-priori probability for class c, respectively. That is, a new sample (pixel) $\mathbf{y}^*$ will be classified to the class c giving the highest value of $g_c(\mathbf{y}^*)$.

In practice, neither the mean vectors nor the covariance matrices are available and hence they have to be estimated from the (very limited) amount of available data. When there are no constraints on the estimates, the maximum likelihood solutions for the means and covariance matrices are the sample means and the sample covariance matrices, denoted by $\tilde{\mu}_c$ and $\tilde{\Sigma}_c$, respectively:

$$\tilde{\mu}_c = \frac{1}{N_c}\sum_{i=1}^{N_c}\mathbf{y}_i, \tag{2}$$

$$\tilde{\Sigma}_c = \frac{1}{N_c}\sum_{i=1}^{N_c}(\mathbf{y}_i-\tilde{\mu}_c)(\mathbf{y}_i-\tilde{\mu}_c)', \tag{3}$$

where N_c is the number of samples in class c. In this paper we will not be altering the mean-value estimates, but focus on putting restraints on the harder-to-estimate covariance matrices.

2.2 Primary and Secondary Subspaces

The models that we focus on are based on separating the feature space into orthogonal *primary* and a *secondary* subspaces. The primary subspace is meant to capture the essential within-class variance, and the "richness" in the probability density modeling within this space is retained, while the secondary space, containing the exterior-class variance, is modeled using a spherical pdf.

To be more specific, let m be the total number of features in the full space (number of spectral bands) and let m_p and m_s be the dimensionality of the primary and secondary spaces, respectively, making $m = m_p + m_s$. Now, letting P_c be the projection matrix for the primary space for class c and $P_{\perp c} = I - P_c$ the corresponding matrix projecting onto the secondary space, we form new covariance matrices like this:

$$\begin{aligned}\hat{\Sigma}_c &= P_c\tilde{\Sigma}P_c + \alpha_c P_{\perp c} I P_{\perp c} \\ &= P_c\tilde{\Sigma}P_c + \alpha_c P_{\perp c},\end{aligned} \tag{4}$$

where the constant α_c is set to $\frac{1}{m_s}\text{tr}\{P_{\perp c}\tilde{\Sigma}_c P_{\perp c}\}$ to ensure that the overall variance is retained, i.e., $\text{tr}\{\hat{\Sigma}_c\} = \text{tr}\{\tilde{\Sigma}_c\}$. Although put a bit loosely, we can say that the variance in the primary space is retained, while the remaining variance is spread spherically, or uniformly, in the secondary space. These covariance estimates, together with the sample means, $\tilde{\mu}_c$, are plugged into the discriminant functions in (1) to form the classifier.

2.3 Subspace Modeling Using PCA

The original SIMCA approach and its derivatives form the primary-secondary space separation by eigenvalue decompositions of the sample covariance matrices. That is, letting V_p^{pca} be an $m \times m_p$ matrix containing the m_p eigenvectors corresponding to the highest eigenvalues and V_s^{pca} the rest of the eigenvectors of $\tilde{\Sigma}_c$ in (3), we choose $P_c = V_p^{pca} V_p^{pca\prime}$. When we use such a P_c, we of course have $\tilde{\Sigma}_c = P_c\tilde{\Sigma}_c P_c + P_{\perp c}\tilde{\Sigma}_c P_{\perp c}$, meaning that there is no (sample) variance across the two spaces. This equality will generally not hold for other choices of P_c.

A slightly different model can be achieved if we do not subtract the class mean when we calculate the sum of outer products constituting $\tilde{\Sigma}_c$, and then eigenvalue-decompose that matrix to get the P_c projection matrix, i.e., we set $\tilde{\mu}_c = 0$ and eigenvalue-decompose (3). By doing this, we include some of the information that is found in the class' mean value when deciding on the primary, or within-class, subspace. When we form the primary space this way, we have $\tilde{\Sigma}_c \neq P_c\tilde{\Sigma}_c P_c + P_{\perp c}\tilde{\Sigma}_c P_{\perp c}$ since there is actual variance across the subspaces which we by our modeling enforce to be zero.

A related approach is that of choosing exactly the same projection matrix for every class. Again we can use the eigenvalue decomposition, although one would make use of the total scatter matrix. This is very similar to the general PCA approach used as a dimensionality reducer, although in (4) we keep the secondary space, i.e., we retain the full dimensionality. Setting $\alpha_c = 0$ for all classes c in (4), on the other hand, would reduce it to the classical PCA approach if the zero-valued eigenvalues were ignored (cf. use of pseudo inverses) in the discriminant function (1).

2.4 Subspace Modeling Using Wider Spectral Bands

In trying to include more domain-specific information into the reduction of model complexity, we propose to form the primary and secondary space separation based on a linear dimensionality reduction technique specifically designed for (continuous) spectral data. Stated more explicitly, we propose to use the dimensionality reduction approach described in [5], which finds the optimal, in a squared-error representation sense, cuts of the spectral curves when forming fewer, but wider, spectral bands. An example of a spectral curve represented with different numbers of wider spectral bands can be seen in Figure 1. Now, by running the algorithm separately with sample sets from the different classes, we get a different set of spectral bands for each class, i.e., each class has its own

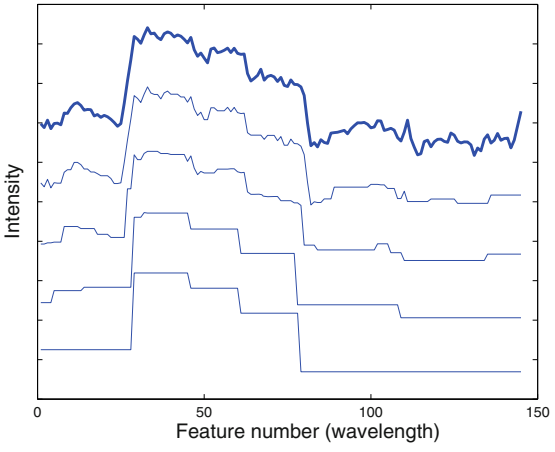

Fig. 1. An example of a spectral curve from the KSC data set represented using different numbers of segments, i.e., using different numbers of primary-subspace dimensions. From top to bottom; using 5, 10, 37, 73 and 145 (maximum, shown in bold) dimensions. The curves are vertically shifted for visual clarity.

linear subspace representing as much of the variance as possible. The idea is that the data-specific restrictions placed on how we form the within-class subspaces make them more generalizable.

Let us say that V is the $m \times m_p$ matrix transforming the data from m to m_p dimensions, or from m to m_p spectral bands, that one obtains using the training samples of a certain class c as input to the just mentioned algorithm. We then obtain the matrix projecting onto the primary space by $P_c = V(V'V)^{-1}V'$, which is then used in (4) to obtain the reduced-complexity covariance estimate used in the discriminant function (1).

Of course, analogous to keeping or ignoring the per-class mean values when we do an eigenvalue decomposition to form the primary subspaces, we can also choose whether or not to remove the per-class means before we run the dimensionality reduction algorithm that finds the new, broader spectral bands. Furthermore, we can use a common projection matrix for all our classes, again analogous to that of the eigendecomposition case described in section 2.2.

2.5 Primary-Subspace Size

What is left to be decided is the dimensionality of the primary subspace for each class. There are several criteria that could be deployed, but in our case we focus on rather small numbers of training samples, and we want to minimize the number of free parameters, hence we choose an equal size for all the classes' primary subspaces. In our experiments, this shared primary-space dimensionality is chosen through crossvalidation on the classification error.

3 Experiments

3.1 Data Sets

To evaluate the classification performance of the discussed approaches we have performed experiments on four hyperspectral images of various sceneries captured using different sensors.

The first image, Pavia [3], is of an urban scene taken by an airborne sensor. It has 71 bands, a pixel size of 2.6 m and the ground truth consists of nine classes. The second image, DC Mall [6], is again from an airborne sensor and also contains urban type data. It is divided into five classes, has a pixel size of 3 m, and has 150 bands. The third and fourth images, Botswana and Kennedy Space Center (KSC), are intended for vegetation inventory [4]. The former was captured by the Hyperion sensor aboard the NASA EO-1 satellite over the Okavango Delta, Botswana on May 31st, 2001. The image has a pixel size of 30 m, the labeled data consists of 14 classes, and the number of raw radiance bands used is 145. The latter data set was captured by an airborne sensor over Kennedy Space Center (KSC) at Cape Canaveral, Florida on March 23rd, 1996. It has a pixel size of about 20 m, has 171 bands, and the ground truth consists of 13 classes. All the above data sets are well known, and the listed references are publications where the data sets are used with various classification algorithms.

3.2 Experiment Details

We compare the classification performance of the normal distribution-based classifier (1) using the covariance matrix estimate of (4) with the six different choices of choosing the primary space projection matrix, P, discussed in section 2. That is, there are three projection matrix schemes found using PCA; not subtracting the class means before calculating the scatter matrices, subtracting the class-means first, and having a common projection matrix for all the classes based on the full scatter matrix, and we have the three proposed corresponding projection matrices using the dimensionality reduction transforms that is based on forming fewer and wider spectral bands.

Furthermore, we also report results using the PCA and the other approach that seeks wider spectral bands to reduce the number of dimensions explicitly, i.e., the same as having a common projection matrix for the classes and ignoring the secondary space completely.

For each of the discussed approaches, the number of dimensions, shared by all classes, of the primary spaces is found using tenfold crossvalidation on the classification error rate. We define the error rate to be the average of the classes' individual error rates.

Each data set was divided into two equally sized, spatially separate, and mostly disjoint training and test sets. We are interested in how the performance varies with training sample size, and hence we have chosen to report results where the total number of training samples are 0.5, 1, 2, 4 and 8 times the dimensionality of each data set. For each training sample size, the experiments

were repeated five times, randomly drawing the selected number of samples from the training set. All data sets were normalized by subtracting the total mean and rescaling the mean within-class variances to one before fitting the models and doing the classification.

3.3 Results

Tables 1 to 4 show the average error rates over the five experiments run for the different training set sizes and for the different data sets. The numbers in parenthesis are the numbers of times the particular classifier gave a lower classification error on the test data than did the other corresponding classifier.

From these numbers we can see that, as expected, the classification error generally decreases with an increased number of training samples. More interestingly, when comparing the proposed way of finding subspaces with that of PCA, we see that the proposed approach seems to dominate at least when modeling each class separately. In the case of a common primary-secondary space separation for all the classes, the proposed approach still performs better than the traditional approach, although not quite as dominantly. When ignoring the secondary spaces, i.e., performing dimensionality reduction, there is no significant difference between the two approaches.

For all data sets, there is a noticeable difference between the results we get when we keep the full dimension of the space and the ones we get when we do a dimensionality reduction. When the number of training samples is very low, about equal to, or lower than, the number of original spectral bands, keeping the full-dimensional feature space gives better results than when doing a dimension reduction. When increasing the number of training samples, the error rates of the two techniques approach each other.

In Figure 2 we show error-rate curves from the Botswana data set when changing the number of dimensions in the primary space. In both the very low training-sample case and where there are quite some more training samples available, we see that there are wider intervals of primary space dimensions that gives acceptable classification errors when applying the proposed approach to finding subspaces. Similar results (not shown) are found using the other data sets.

3.4 Discussion

The way that the proposed approach finds the subspaces is much more restricted than that using PCA, as the proposed approach is based on finding a linear basis that consists of the average of contiguous original spectral bands instead of allowing any linear combination of the original bands. The results seem to indicate that we are capable of modeling the within-class variance properly, while avoiding overfitting the training data.

The dimensionality of the primary space is found through crossvalidation and it seems like a lot of the success of the proposed approach stems from the wider range of such primary-space dimensions that give acceptable results. That is, there is a greater number of choices of primary-space dimensions that give good

Table 1. Mean classification errors on the 5 experiments per training set size using the DC Mall data set. On the left of the slash we show results based on the proposed way to find subspaces, on the right the traditional PCA-based one. In parenthesis one can find the number of "wins" (out of the 5 repeated experiments) for the classifier.

Train-set size	Mean included	Mean excluded	Common P	Dim. reduction
0.5	13.7 (5) / 25.1 (0)	14.9 (5) / 26.4 (0)	14.1 (3) / 15.6 (2)	20.7 (0) / 18.2 (5)
1.0	12.5 (5) / 26.8 (0)	10.6 (5) / 23.5 (0)	10.0 (5) / 12.0 (0)	17.2 (3) / 15.3 (2)
1.5	9.5 (5) / 22.0 (0)	9.4 (5) / 19.7 (0)	9.7 (5) / 11.0 (0)	11.1 (4) / 11.3 (1)
2.0	11.4 (5) / 19.0 (0)	9.8 (5) / 19.6 (0)	10.6 (4) / 13.9 (1)	13.3 (2) / 13.2 (3)
4.0	13.0 (5) / 20.7 (0)	10.3 (5) / 20.5 (0)	9.5 (3) / 10.5 (2)	10.3 (1) / 8.6 (4)
8.0	8.9 (5) / 16.3 (0)	8.9 (5) / 15.9 (0)	9.2 (2) / 9.5 (3)	7.8 (4) / 8.5 (1)

Table 2. Data set: Pavia. For explanation, see the caption of Table 1.

Train-set size	Mean included	Mean excluded	Common P	Dim. reduction
0.5	24.7 (5) / 31.8 (0)	23.6 (4) / 32.4 (1)	29.3 (2) / 30.2 (3)	54.2 (2) / 50.8 (3)
1.0	16.6 (4) / 22.3 (1)	22.6 (2) / 20.1 (3)	15.2 (5) / 24.1 (0)	28.3 (2) / 23.8 (3)
1.5	11.7 (5) / 22.2 (0)	13.2 (5) / 17.7 (0)	12.1 (5) / 21.2 (0)	20.9 (2) / 19.9 (3)
2.0	11.0 (5) / 16.3 (0)	12.4 (5) / 16.1 (0)	9.6 (5) / 15.2 (0)	15.7 (2) / 15.7 (3)
4.0	9.4 (4) / 21.6 (1)	11.3 (4) / 13.1 (1)	10.1 (5) / 13.5 (0)	12.7 (5) / 13.6 (0)
8.0	8.1 (5) / 11.8 (0)	9.0 (5) / 12.7 (0)	7.0 (5) / 11.0 (0)	10.3 (4) / 11.6 (1)

Table 3. Data set: KSC. For explanation, see the caption of Table 1.

Train-set size	Mean included	Mean excluded	Common P	Dim. reduction
0.5	34.4 (5) / 42.9 (0)	37.8 (5) / 92.3 (0)	32.8 (4) / 32.4 (1)	62.3 (3) / 63.1 (2)
1.0	28.2 (4) / 30.4 (1)	26.7 (5) / 30.0 (0)	29.1 (2) / 30.0 (3)	34.8 (1) / 34.8 (4)
1.5	23.6 (5) / 30.5 (0)	21.4 (5) / 28.5 (0)	22.6 (5) / 26.1 (0)	29.4 (1) / 26.3 (4)
2.0	21.3 (5) / 29.6 (0)	19.6 (5) / 25.4 (0)	20.7 (5) / 24.8 (0)	23.5 (2) / 23.5 (3)
4.0	20.0 (5) / 29.6 (0)	19.2 (5) / 26.1 (0)	19.1 (2) / 18.7 (3)	19.0 (2) / 19.2 (3)
8.0	17.0 (5) / 29.8 (0)	16.7 (5) / 27.6 (0)	16.1 (0) / 14.7 (5)	15.4 (3) / 16.1 (2)

Table 4. Data set: Botswana. For explanation, see the caption of Table 1.

Train-set size	Mean included	Mean excluded	Common P	Dim. reduction
0.5	25.2 (5) / 32.1 (0)	22.7 (5) / 33.0 (0)	25.5 (3) / 28.4 (2)	38.8 (5) / 41.5 (0)
1.0	15.8 (5) / 28.8 (0)	16.0 (4) / 20.6 (1)	14.2 (5) / 23.0 (0)	22.8 (2) / 22.1 (3)
1.5	14.9 (3) / 19.9 (2)	13.2 (5) / 21.1 (0)	16.8 (4) / 18.3 (1)	19.4 (2) / 19.1 (3)
2.0	14.3 (4) / 17.6 (1)	10.8 (5) / 18.7 (0)	11.1 (4) / 13.0 (1)	15.6 (0) / 13.5 (5)
4.0	9.9 (5) / 18.7 (0)	10.1 (5) / 17.7 (0)	10.4 (2) / 9.2 (3)	8.9 (3) / 9.1 (2)
8.0	8.9 (5) / 19.1 (0)	8.3 (5) / 17.4 (0)	7.6 (4) / 8.2 (1)	7.8 (3) / 8.8 (2)

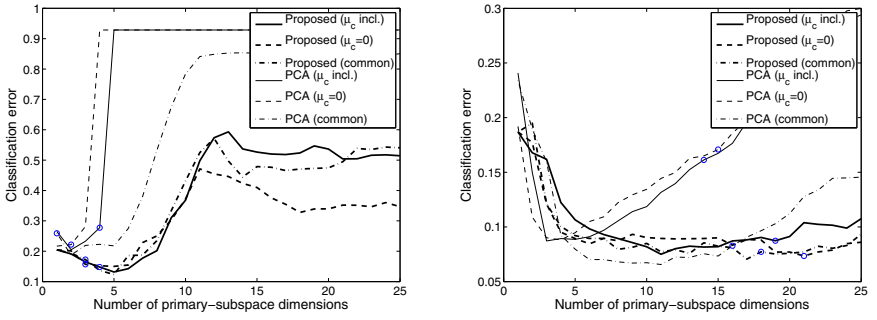

Fig. 2. Classification error rates when varying the size of the primary subspace for the Botswana data set. Training set sizes are (a) about equal to the number of original spectral bands and (b) about 8 times as many. Note how quickly the error rates for the PCA-based approaches increase with added dimensions, making it hard for any crossvalidation technique to choose a good primary-space size. Subspace sizes chosen by crossvalidation for these particular training sets are marked by 'o's.

classification results using the proposed approach than there are choices giving good results using PCA. After adding a few PCA-dimensions to the primary space, the whole sample variance is captured, leaving us with an artificial, or overfitted, primary space, while at the same time there is no variance left to "spread out" over the secondary space.

Especially in the case of a very limited set of training samples, it seems to be a good idea to avoid doing a "crisp" dimensionality reduction, but rather keep the secondary space, although with a simpler pdf model. When there are very few training samples, it is important to try to keep as much as possible of the space that they span, while at the same time avoid overfitting. Modeling the "surplus" space using a simpler model, rather then ignoring it, seems to be a good compromise.

4 Conclusion

Modeling the covariance is a key factor in normal distribution-based classifiers. When there is a need to restrict the complexity of such models, one rather flexible approach is to form orthogonal subspaces of the feature space, and let the variance in each of them be modeled with a different complexity. In this paper we have studied the classifier performance on hyperspectral image data when applying different approaches to forming these subspaces. In particular, we have proposed to use subspaces created by forming fewer and wider spectral bands instead of the more general PCA. The results indicate that the classifiers seem to benefit from using this more data-specific approach.

References

1. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley Interscience, Hoboken (2000)
2. Frank, I.E.: Dasco: a new classification method. Chemometrics and Intelligent Laboratory Systems 4(3), 215–222 (1988)
3. Gamba, P.: A collection of data for urban area characterization. In: Proc. IEEE Geoscience and Remote Sensing Symposium (IGARSS 2004), pp. 69–72 (2004)
4. Ham, J., Chen, Y., Crawford, M.M., Ghosh, J.: Investigation of the random forest framework for classification of hyperspectral data. IEEE Trans. Geosci. Remote Sensing 43(3), 492–501 (2005)
5. Jensen, A.C., Solberg, A.S.: Fast hyperspectral feature reduction using piecewise constant function approximations. IEEE Geoscience and Remote Sensing Letters 4(4), 547–551 (2007)
6. Landgrebe, D.A.: Signal Theory Methods in Multispectral Remote Sensing. Wiley Interscience, Hoboken (2003)
7. Næs, T., Indahl, U.: A unified description of classical classification methods for multicollinear data. Journal of chemometrics 12(3), 205–220 (1998)
8. Wold, S.: Pattern recognition by means of disjoint principal components models. Pattern Recognition 8(3), 127–139 (1976)