Improving Hyperspectral Classifiers: The Difference Between Reducing Data Dimensionality and Reducing Classifier Parameter Complexity

Asbjørn Berge and Anne Schistad Solberg

Department of Informatics
University of Oslo, Norway
asbjorb@ifi.uio.no

Abstract. Hyperspectral data is usually high dimensional, and there is often a scarcity of available ground truth pixels . Thus the task of applying even a simple classifier such as the Gaussian Maximum Likelihood (GML) classifier usually forces the analyst to reduce the complexity of the implicit parameter estimation task. For decades, the common perception in the literature has been that the solution to this has been to reduce data dimensionality. However, as can be seen from a result by Cover [1], reducing dimensionality increases the risk of making the classification problem more complex.Using the simple GML classifier we compare state of the art dimensionality reduction strategies with a recently proposed strategy for sparsing of parameter estimates in full dimension [2]. Results show that reducing parameter estimation complexity by fitting sparse models in full dimension have a slight edge on the common approaches.

1 Introduction

Hyperspectral imaging, an increasingly common tool in remote sensing, is sampling of the spectrum of reflected sunlight in wavelengths from ultraviolet to infrared. As a natural extension of the multispectral sensors, hyperspectral sensors sample reflected sunlight in 50 to several hundred contiguous narrow bands. Thus more information can be extracted from a single pixel compared to a multispectral image, however the high dimension of the resulting feature space makes classification of pixels a complex problem. Features also usually exhibit high correlation, adding a redundancy to the data that in some cases may obscure the information important for classification. When the number of training samples is low compared to data dimensionality the so called curse of dimensionality impacts the generalization capability of the classifiers designed.

The common approach for dealing with the curse of dimensionality in the literature is to reduce the dimensionality, and thus indirectly reducing the number of parameters to estimate. Contrary to this approach, it is possible to reduce the number of parameters to estimate by choosing to fit simpler models in full

B.K. Ersbøll and K.S. Pedersen (Eds.): SCIA 2007, LNCS 4522, pp. 293–302, 2007.

dimension. We will discuss the simple classifier resulting from Bayes rule when assuming that classes are distributed as Gaussians. The main contribution of this paper is to present results and a discussion comparing indirect (dimensionality reduction) and direct (parameter sparsing) simplifications of such classifiers. The motivation for this comparison can be found in Covers theorem [1]. We want to ascertain whether dimensionality reduction can be seen to make the classification problem more complex.

In section 2 and section 3 we present the classifier, and point out the effect of Covers theorem. Section 4 discusses the contrary approach of reducing the number of parameters to estimate by fitting a sparse model. In section 5 we briefly review some of the dimensionality reduction strategies proposed in the literature on classification on remotely sensed hyperspectral data. Section 6 presents and discuss the results of several experiments on four different hyperspectral images. Section 7 concludes this paper.

2 The Classification Task

Consider a classification problem with k classes, assuming class conditional distributions to be Gaussian with mean μ_k and class-wise covariance matrices Σ_k. It is well known that this reduces to comparing the k quadratic discriminant functions $g_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k) + \log\pi_k$, where π_k is the a priori probability for class k. The parameters of these distributions are usually calculated from the maximum likelihood estimates and plugged into the above rule. This decision rule is commonly referred to a Gaussian Maximum Likelihood(GML) classifier. The common problem with the GML classifier is that the number of parameters to estimate grows quadratically with the dimensionality of the feature space. Clearly this means that we quickly run out of samples to reliably estimate these parameters.

3 The Separability of Patterns as a Function of Dimensionality

Our classification problem may be an intrinsically non-linear problem, or may *become* a non-linear problem after dimensionality reduction according to Cover's theorem [1] regarding the separability of patterns. For a set of N samples, and a classifier represented by a surface with d degrees of freedom, where any labeling is equally probable, the probability of randomly picking a class labeling of samples that can be perfectly separated by the chosen classifier is [1]

$$P(N,d) = (\frac{1}{2})^{(N-1)} \sum_{k=0}^{d-1} \binom{N-1}{k}.$$

Plainly what this states is that by increasing the degrees of freedom of a surface intended to separate the classes, the probability of being able to separate the

classes approaches one. This is a very common justification for the efficacy of support vector machines and other kernel methods, where you map the data up into a much higher dimensional space and solve a linear classification problem. A linear decision boundary between two classes in a space with m features has $d = m+1$ degrees of freedom corresponding to the dimensionality m of its normal vector, plus one allowing for an arbitrary intercept. Likewise, fitting a quadratic decision boundary corresponds to $d = \binom{m+2}{2}$ degrees of freedom, i.e., finding the coefficients and intercept of a quadratic function in the feature space. When introducing a dimensionality reduction on the feature space, this would be the same as only allowing decision boundaries in some subspace, so consequently the degrees of freedom will be reduced as a function of the reduced dimensionality. Thus, for a fixed number of samples, reducing the dimensionality by any feature extraction or selection method reduces the probability of randomly picking a labeling that can be separated - and for a fixed number of samples this probability is dependent on the degrees of freedom of the classifier. Consequently, it might be harder in reduced dimensions to find a linear classifier that separates the data than a more complex classifier. In section 6 we compare the classification performance of a linear and quadratic classifier as a function of dimensionality to observe this phenomenon with real data.

4 Parameter Sparsing in Full Dimension

Crude models for reducing the number of parameters to estimate in a GML classifier are well known. Constraints such as the assumption that features are uncorrelated, well known as a naïve bayes classifier, reduces the number of parameters to estimate for each distribution down to the dimensionality. We recently [2] proposed an approach for reducing the number of parameters needed to estimate when designing classifiers in high dimensional feature spaces, sparse cholesky triangle inverse covariance (STIC) estimates. The method is based on time series theory regarding the Cholesky decomposition of the inverse covariance matrix, $\Sigma_k^{-1} = L_k D_k L_k^T$. where L_i is a lower triangular matrix with ones on the diagonal and D a diagonal matrix. (See Fig. 1a) If we were to consider the features of each sample as a time-series, the elements in L can be seen rowwise as parameters in autoregressive processes of the same order as the row r. Several authors in the time series literature have noted this [3], [4]. We will use this fact to transform the task of approximating covariance matrices into a sequence of regressions. For each row, r, one could then "predict" the next feature based on the $r-1$ preceding features. Assuming zero mean for readability, this can be expressed as: $x_r = \sum_{j=1}^{r-1} \alpha_{r,j} x_j + \varepsilon_r$ where the rth diagonal entry of $D_{r,r} = \mathrm{var}(\varepsilon_r)$. This parametrization has the effect that the resulting covariance matrix will still be positive definite, as long as the diagonal elements of D are positive.

The general idea is to start by approximating the covariance matrices with the simplest possible models, i.e.,diagonal matrices, and add parameters to the

$$L = \begin{bmatrix} 1 & & & & \\ -\alpha_{2,1} & 1 & & & \\ -\alpha_{3,1} & -\alpha_{3,2} & 1 & & \\ \vdots & & & \ddots & \\ -\alpha_{p,1} & & -\alpha_{p,2} & -\alpha_{p,p-1} & 1 \end{bmatrix}$$

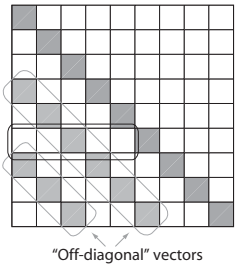


Fig. 1. Illustration of a matrix of correlations, L, for the inverse covariance matrix. The matrix is lower triangular, with 1 on the diagonal. Sparsity in the covariance estimate is obtained by only estimating the matrix elements in *some* off-diagonal vectors. The matrix is estimated by a sequence of regressions, one for each row in the matrix L, indicated with black rectangle.

approximation until the classification performance of the model no longer improves. A very simple heuristic, illustrated in Fig. 1b is to gradually add off-diagonal elements one diagonal vector at a time in a sequential forward fashion. One specific diagonal vector corresponds to the set of correlations with some specific distance between features. The search is guided by ten-fold cross-validation (10-CV) as a performance measure. Further details of this approach can be found in [2].

5 Dimensionality Reduction

In the literature on classification of remotely sensed hyperspectral data several linear feature extraction techniques has been proposed with the goal of improving the performance of the classification. In our comparison we include two classic approaches and two approaches designed for hyperspectral data, decision boundary feature extraction (DBFE) [5] and non-parametric weighted feature extraction (NWFE) [6].

By far the most common dimensionality reduction method in multivariate statistics is the well known principal component analysis (PCA) which eigenanalyze the scatter (or equivalently covariance) matrix of the dataset, corresponding to projecting the data onto a set of orthogonal vectors where the scatter (variance) of the data is highest. Another classical feature extraction technique is Fisher's linear discriminant (LDA).

A method for eigenanalyzing the "scatter" of the decision boundary, called decision boundary feature extraction (DBFE), was proposed in [5], with applications on hyperspectral data in mind. DBFE has been relatively popular in the literature on classification of hyperspectral images. The general idea is to find a set of virtual samples as intersection points between the decision boundary and lines between samples of different classes. The principal components of these virtual samples can thus be argued to describe the most important vectors in the decision boundary and possibly discriminative features. The heuristic used

for finding these virtual samples is to classify the dataset in full dimension and then intersecting the estimated decision boundary with lines between the closest samples in opposite classes.

Another popular feature extraction method developed for hyperspectral images is the nonparametric weighted feature extraction(NWFE) proposed in [6], which is a nonparametric extension of LDA by redefining the scatter according to distance. By weighing the influence of samples according to the distance to samples in opposite classes, samples near the decision boundary are considered more important. In LDA, the between-class scatter is of deficit rank. In NWFE this is overcome by redefining this scatter to represent the scatter of between samples and a distance-weighted mean in an opposite class.

In the literature, several feature selection approaches have been proposed, tailored for specific hyperspectral classification tasks. However, for our purposes it is reasonable to compare with feature selection methods that does not need any initialization of the number of features wanted, and keeps the features in the set after selection. Reasonably, when a modest number of data is available for training, the optimal number of features might be low. Sequential forward search (SFS) is the simplest possible algorithm in such cases, adding features sequentially ranked by some criterion, the experiments presented here use the Mahalanobis distance.

6 Experimental Results and Discussion

Four hyperspectral datasets are analyzed in this work. The first, *Fontainebleu*, is from an airborne sensor (ROSIS), containing forest pixels from Fontainebleu south of Paris. It is divided into three classes, have 81 bands and a pixel size of 5.6m. The second dataset *Pavia*[7] is also from an airborne sensor (DAIS), depicting urban landcover pixels over Pavia, Italy. The dataset has 71 bands and 2.6m pixels. The third dataset contains a wetland vegetation scene acquired over the Okavango delta in *Botswana*[8] acquired by the Hyperion sensor aboard the EO-1 satellite, with a total of 145 bands after removal of uncalibrated and noisy bands. The image has 30m pixels. The last image we use is the from an AVIRIS airborne sensor, over Kennedy Space Center (*KSC*)[8], is a vegetation dataset, with 18m pixels and 176 bands. For all the datasets, the average number of training pixels per class is 700, 100, 196, and 115 in presented order. For all datasets we designed (as far as it was possible) spatially separate datasets for training and testing to avoid fitting the classifiers to the similarities between neighboring pixels due to spatial correlation. The same set of 10-fold cross-validation rotation on the training data was used for model choice in all methods, i.e., guiding the number of features or the number of nonzero parameters. The reported performance is average overall correct classification rate. We compare classification performance for the models chosen by cross-validation of the different dimensionality reduction methods: principal component analysis (PCA), Fisher's linear discriminant (LDA), non-parametric weighted feature extraction (NWFE), decision boundary feature extraction (DBFE), sequential forward

feature selection (SFS) and the parameter sparsing approach using sparse cholesky triangle covariance estimates (STIC).

6.1 Results

Table 1 reports the test set classification performance and the number of parameters estimated (as a fraction of a full dimensional GML classifier). The reported results are included as a supplement to the results given in the figures. One notes that STIC has a slight edge on all dimension reduction strategies, however, an interesting result is that we commonly use more degrees of freedom to estimate sparse models in full dimension.

In Fig. 2a-2d the performance of a linear classifier versus a quadratic classifier as a function of the number of features (extracted by PCA) is given. The solid black line in these plots indicate the performance of the chosen STIC model. Several conclusions can be made from the presented results. We can observe

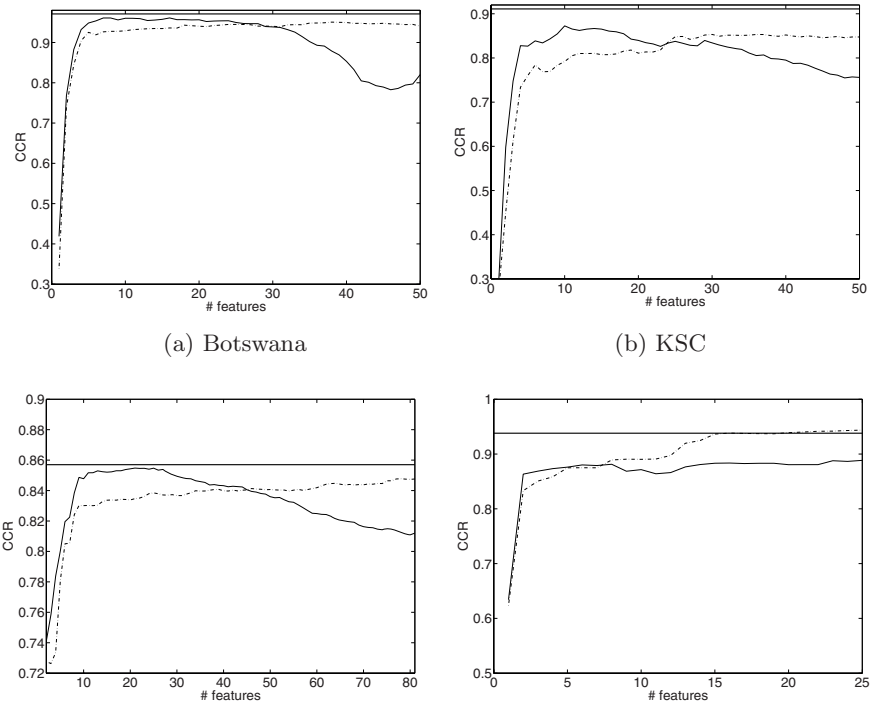
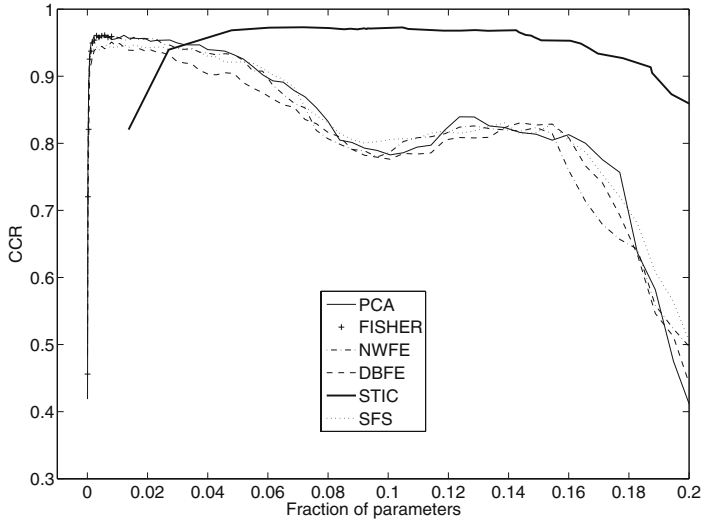
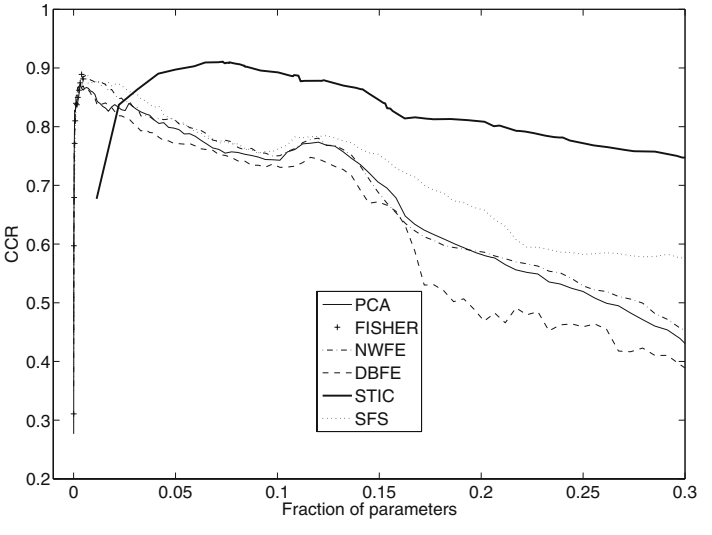


Fig. 2. Comparison of correct classification on test data to the number of features retained using PCA as dimension reduction. The performance of a linear classifier is represented by a stippled line and a quadratic classifier by a solid line. The thick solid line indicates the performance of the STIC model chosen by cross-validation. For visualization purposes, only the 50 first features is shown for the KSC and Botswana images and the first 25 for the Pavia image.



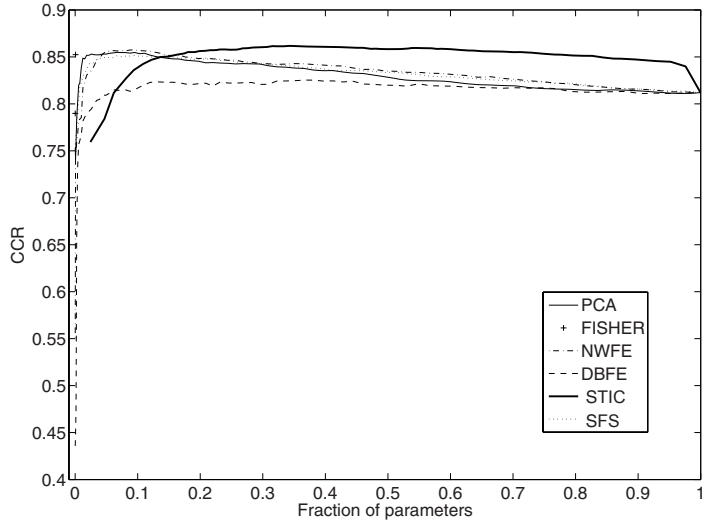
(a) Botswana



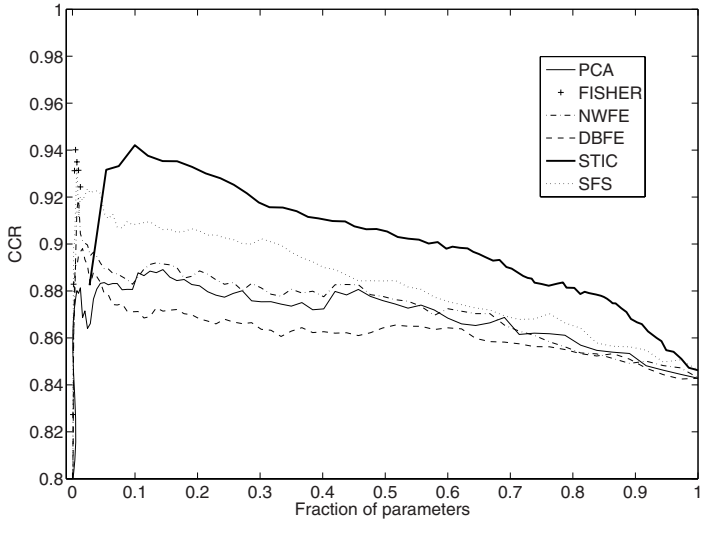
(b) KSC

Fig. 3. Correct classification rates on test data compared to the fraction of covariance parameters for a full model. All methods decay quite rapidly for cases using more than 30% of the parameters of a full model.

similarities in Fig. 2a-2d. All datasets have reasonably low amounts of data available for training. Thus when dimensionality increases, the number of degrees of freedom of the quadratic classifier grows so fast that the estimates become unstable, due to the curse of dimensionality. Not unexpectedly the simpler linear classifier decays slower, and overtakes the quadratic classifier at some point.



(a) Fountainebleu



(b) Pavia

Fig. 4. Correct classification rates on test data compared to the fraction of covariance parameters for a full model

These results can also be viewed as support for the conclusions drawn from Cover's theorem, arguing that when dimension is low, the decision boundary tends to be less linear than it is in full dimension. This indicates that parameter sparsing by STIC, i.e., a search for simple classifier models in full dimension is reasonable.

Table 1. Average overall test set performance for the models chosen by cross-validation for the different datasets. In parentheses the number of parameters estimated as the fraction of a full model. All measures in percent.

Dataset	Fontainebleu	Pavia	Botswana	KSC
PCA	85.2(4.8)	87.5(41.8)	95.8(1.1)	86.7(0.6)
LDA	85.2(0.06)	93.4(0.85)	95.8(0.7)	88.9(0.4)
NWFE	85.7(7.7)	88.4(35.8)	95.4(1.4)	88.6(0.5)
DBFE	82.5(52.3)	87.1(13.2)	93.8(1.9)	86.7(0.4)
SFS	85.1(6)	90.7(10.9)	94.7(1.2)	87.5(1.05)
STIC	85.7(23)	93.8(17)	97.1(9.6)	91.1(8.5)

Fig. 3a-4b illustrates the performance for the different dimensionality reduction strategies as a function of the number of parameters estimated compared to a full model. In these plots, the performance of the parameter sparsing strategy, STIC, is also given. The general conclusion that can be drawn from Fig. 3a-4b is that fitting sparse models in full dimensional space is fairly effective over a wide range of parameter sizes. Even so, the best model found for the STIC, especially in the case of the high dimensional images see Fig. 3a and 3b, estimates a lot more parameters than the correspondingly optimal models using dimensionality reduction. (See Table 1.) One can note that the dimensionality reduction results from these images are fairly similar for all feature extraction and selection methods. One possible explanation for this might be that since the features are so highly correlated, any feature reduction will cover mostly the same discriminative information, regardless of approach. From Fig. 4a and 4b we can observe typical performance of full dimensional sparsed models over the entire range. As can be seen, when dimensionality is fairly low, and the amount of training data available is high, as with the Fontainebleu image, little is gained by using sparse models compared with feature extraction. The classes in this dataset is known to be overlapping even in the full dimensional space, and two of the three classes are extremely similar, so this dataset might be complex to classify even in full dimension. Considering class separability, the Pavia image is an example of the opposite - classes can be reasonably classified using a linear classifier in full dimension. This can be seen in the fairly high performance of LDA as a feature extractor.

7 Conclusion

We have discussed the difference between reducing classifier complexity using dimension reduction versus parameter reduction. Theoretical results [1], and supporting experimental results indicate the soundness of fitting simple models in full dimensional space compared to using more complex classifiers after reducing dimensionality. Specifically, our previously proposed strategy, STIC, seems to have a slight edge on dimensionality reduction. However, STIC is still more of a proof of concept than a fully developed method. The heuristic used for

selection of non-zero parameters is a bit crude, and as can be seen in Table 1, we usually use fairly many degrees of freedom for describing the classifier.

References

1. Cover, T.M.: Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Transactions on Electronic Computers (3), 326–334 (1965)
2. Berge, A., Jensen, A.C., Solberg, A.S.: Sparse inverse covariance estimates for hyperspectral image classification. IEEE Trans. Geosci. Remote Sensing, Accepted for publication (2007)
3. Smith, M., Kohn, R.: Parsimonius covariance matrix estimation for longitudinal data. Journal of the American Statistical Association 97(460), 1141–1153 (2002)
4. Pouhramadi, M.: Foundations of Time Series Analysis and Prediction Theory. Wiley, Chichester (2001)
5. Lee, C., Landgrebe, D.: Feature extraction based on desicion boundaries. IEEE Trans. Pattern Anal. Machine Intell. 15(15), 388–400 (1993)
6. Kuo, B.C., Landgrebe, D.: A robust classification procedure based on mixture classifiers and nonparametric weighted feature extraction. Remote Sensing 40(11), 2486–2494 (2002)
7. Gamba, P.: A collection of data for urban area characterization. In: Proc. IEEE Geoscience and Remote Sensing Symposium (IGARSS'04) (2004)
8. Ham, J., Chen, Y., Crawford, M.M., Ghosh, J.: Investiagtion of the random forest framework for classification of hyperspectral data. Remote Sensing 43(3), 492–501 (2005)