Regression Approaches to Small Sample Inverse Covariance Matrix Estimation for Hyperspectral Image Classification

Are C. Jensen, Student Member, IEEE, Asbjørn Berge, Member, IEEE, and Anne Schistad Solberg, Member, IEEE

Abstract—A key component in most parametric classifiers is the estimation of an inverse covariance matrix. In hyperspectral images, the number of bands can be in the hundreds, leading to covariance matrices having tens of thousands of elements. Lately, the use of linear regression in estimating the inverse covariance matrix has been introduced in the time-series literature. This paper adopts and expands these ideas to ill-posed hyperspectral image classification problems. The results indicate that at least some of the approaches can give a lower classification error than traditional methods such as the linear discriminant analysis and the regularized discriminant analysis. Furthermore, the results show that, contrary to earlier beliefs, estimating long-range dependencies between bands appears necessary to build an effective hyperspectral classifier and that the high correlations between neighboring bands seem to allow differing sparsity configurations of the inverse covariance matrix to obtain similar classification results.

Index Terms—Cholesky decomposition, covariance parameterization, hyperspectral image classification, pattern classification, precision matrix, regularization, sparse regression.

I. INTRODUCTION

HYPER SPECTRAL image classification operates, by definition, on data of high dimensionality, and the number of training pixels is often severely limited. Furthermore, the high correlation between features adds a redundancy to the data that may sometimes obscure the information important for classification.

When using parametric methods, such as the Gaussian maximum likelihood classifier, the parameter estimates, most importantly the covariance matrix estimate, will become increasingly unstable when the number of labeled samples is low compared to the dimensionality of the feature space. This instability quickly deteriorates the generalization ability of the classifier. In dealing with this problem, one can try to reduce the dimensionality of the feature space and/or bias the parameter estimates toward simpler and more stable estimates. This paper focuses solely on the latter using extensions of well-known time-series theory to reduce the task of regularizing covariance matrix estimates to regularizing sequences of linear regressions.

A traditional way of stabilizing the covariance estimate and, hence, increasing the generalization performance is to replace the sample covariance matrix \( \hat{\Sigma} \) by

\[
\hat{\Sigma} = (1 - \alpha) \hat{\Sigma} + \alpha I
\]

where \( I \) is a scaled identity matrix or the diagonal of \( \hat{\Sigma} \) and \( \alpha \) is a parameter controlling the amount of regularization. Using cross-validation estimates of the classification performance to determine \( \alpha \), (1) can be seen as a variant of regularized discriminant analysis (RDA) [1], and by using leave-one-out cross-validation on the likelihood, one gets a simplified leave-one-out covariance matrix estimate (LOOC) [2].

The concept of using linear regressions in estimating the inverse covariance matrix comes from the modified Cholesky decomposition \( \Sigma^{-1} = T' DT \), in which each row of the lower triangular \( T \) can be interpreted as coefficients of a linear regression. This has been well known for several decades in the time-series literature [3]. By the use of coefficient shrinkage and subset selection, these regressions can be made to handle small sample numbers and, in consequence, provide a more robust estimate of the inverse covariance matrix.

In the time-series literature, particularly considering regression analysis of longitudinal time series (in which “ordered” measurements are performed for the objects under analysis), the application of linear regressions as a tool for modeling large covariance matrices has seen a resurgence in the later years. The strategies proposed range widely, from imposing a band-structure [4], [5] (antedependence modeling) on the inverse covariance matrix, extended by assuming a polynomial structure on the antedependence model [6] or by smoothing of the lower triangular Cholesky factors [7]. Penalized regression has also been applied to estimate shrunken or sparsed Cholesky factors [8], [9].

As pointed out, most of these papers were concerned with covariance estimates in regressions intended for forecasting longitudinal time series. There is clearly a need to investigate how these and other well-known regression techniques perform in estimating large covariance matrices in a classification setting.

The statistical toolbox for regression when the predictors are highly correlated, as is the case with hyperspectral data, contains several well-established approaches. Ridge regression...
and principal component regression (PCR) are examples of regularization strategies, with the effect of biasing the regression solutions away from directions where the predictors have low spread [10]. A generalization of ridge regression, the least absolute shrinkage and selection operator (LASSO) [11], replaces the $L_2$ ridge penalty by the $L_1$ penalty, thereby performing a kind of continuous feature selection. Another branch of approaches for stabilizing the sequences of regressions is predictor selection techniques, like the sequential forward selection (SFS) [12].

Recently, several papers have focused on using support vector machine (SVM) classifiers for the classification of hyperspectral data [13]. There is some dispute about the claim that these methods are insensitive to dimensionality issues and overtraining. SVM classifiers are related to this paper as both approaches search for sparse models for classification. The cost of misclassification of training data is the SVM measure of regularization. The SVM is a so-called kernel method, meaning that it measures sample distance in some space implicitly defined by a weighting function (kernel). A common choice for SVM kernel is the Gaussian radial basis function (RBF), where one applies a Gaussian kernel to each sample, and the “kernel trick” allows one to evaluate the distance measure without mapping the data into a higher dimensional space. This kernel requires the width of the Gaussian kernel to be determined.

It should be mentioned that, although not used in our experiments, unlabeled data can be included to help stabilize both the parametric methods and the SVM [14]–[16].

In this paper, we have adapted general approaches for stabilizing regressions for application in covariance estimation for classification, and results from extremely ill-posed scenarios are reported. To establish performance baselines for the evaluation of the different regression strategies, a conventional RDA and the implicitly regularized SVMs have been applied.

The following section details the linear classifier used in this paper. Section III gives a short description of the parameterization leading to covariance estimation by linear regression. Section IV details known ways of regularizing and structuring these regressions and discusses slight heuristic improvements. Experimental methodology and the data sets are discussed in Section V, and the results of the experiments are presented in Section VI. Section VII discusses the results and concludes this paper.

II. LDA CLASSIFIER

When the number of available training samples is limited, a common classifier is the linear discriminant analysis (LDA) classifier which assumes Gaussian densities with a common covariance matrix for all the classes. Let $y$ be a row vector containing the $p$ hyperspectral band values (features) of a pixel. Applying minimum-error-rate classification yields the following discriminant function for each class $k$ [17]:

$$g_k(y) = -\frac{1}{2}(y - \mu_k)'\Sigma^{-1}(y - \mu_k) + \log \pi_k$$

(2)

where $\mu_k$ and $\pi_k$ are the mean vector and a priori probability for class $k$, respectively, and $\Sigma^{-1}$ is the common inverse covariance matrix for which estimation and regularization is discussed in this paper. Note that only the inverse of the covariance matrix is needed for classification. The discriminant functions (2) are used throughout this paper.

III. REGRESSION APPROACH TO INVERSE COVARIANCE MATRIX ESTIMATION

The following shows how the inverse covariance matrix can be decomposed into a series of regressions. For a more detailed introduction to this well-known procedure, see [8].

Let $y$ be a row vector of band values and assume zero mean for simplicity. Now, define

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

(3)

where the $\alpha$’s are the regression coefficients obtained by regressing each feature on its predecessors, i.e., finding the $\alpha$’s of minimizing the mean squared error.

The regressions remove any linear dependence between the residual elements $\varepsilon = Ty$, thus making its covariance matrix diagonal

$$D = E[\varepsilon\varepsilon'] = E[Tyy'T'] = T\Sigma T'$$

(4)

where $\Sigma$ is the covariance matrix of $y$ that is needed, in its inverse form, for classification (2). By assuming a multivariate normal distribution, the diagonal covariance (4) results in independent residual elements, i.e.,

$$P(y) = \prod_{r=1}^{p} P(\varepsilon_r) = \prod_{r=1}^{p} P(y_r|y_{1:r-1})$$

(5)

which can be used in maximum likelihood (ML) estimation. Note that (4) holds independently of the distribution of $y$.

By rearranging (4), we get

$$\Sigma^{-1} = T'D^{-1}T$$

(6)

which is the modified Cholesky decomposition of $\Sigma^{-1}$. As long as the diagonal elements of $D = \text{diag}(\text{var}(\varepsilon_1), \text{var}(\varepsilon_2), \ldots, \text{var}(\varepsilon_p))$ are positive, any choice of $\alpha$’s will produce a positive definite covariance matrix. Thus, by using different regression approaches and retaining a diagonal $D$, the resulting regression coefficients $\alpha_{r,j}$ can be used to estimate the inverse covariance matrix in (6) for subsequent use in the discriminant function of (2). Note that (2) can be evaluated using $D^{-1}$ and $T$ directly without explicitly calculating $\Sigma^{-1}$.
A. Parameter Estimation

1) Least Squares: Independent of the distribution of \( y \), the estimates of the elements of \( T \) can be found by minimizing the squared residual error

\[
\min_{\alpha_r} \sum_{i=1}^{N} \varepsilon_{i,r}^2, \quad \alpha_r = (\alpha_{r,1}, \alpha_{r,2}, \ldots, \alpha_{r,r-1})
\]  

(7)

where \( N \) is the number of pixels. Equation (7) can be solved using simple linear algebra. \( D \) can be obtained by using the sample estimates of the variances of the residuals.

2) ML: By assuming a multivariate normal distribution, ignoring irrelevant constants, and using \( |\Sigma| = |D| = \prod_{r=1}^{p} \sigma_r^2 \), a log-likelihood based on (5) can be written \[8\]

\[
-l(\Sigma; y_r) \propto \sum_{r=1}^{p} N \log \sigma_r^2 + \sum_{i=1}^{N} \frac{\varepsilon_{i,r}^2}{\sigma_r^2}.
\]

(8)

Maximizing this likelihood with respect to the elements of \( T \) (the \( \alpha \)'s) yields the same solution as minimizing the squared error in (7), i.e., each row of \( T \) or \( r \) in (8) can be solved independently as linear regressions.

IV. REGULARIZING AND STRUCTURING THE REGRESSIONS

A. Regression Shrinkage/Penalized Regression

Several approaches have been developed in the chemometrics literature to reduce the instability of the coefficient estimates in regressions \[10\]. By biasing or shrinking the regression coefficients, the variance in the predictions can be reduced. Two common approaches are ridge regression \[10\] and LASSO \[11\], which both impose a penalty on the size of the regression coefficients. Assuming Gaussian noise, these penalizations are equivalent to placing a priori distributions on the regression coefficients in the ML setting. Penalizing the likelihood in (8) yields \[8\]

\[
-l_{\text{penalized}}(\Sigma; y_r) = -l(\Sigma; y_r) + \lambda \sum_{r=2}^{p} \sum_{j=1}^{r-1} |\alpha_{r,j}|^{\phi}
\]

(9)

where \( \lambda \) is the tuning parameter and \( \phi = 2 \) gives an equivalent to using a Gaussian prior, known as ridge regression, and \( \phi = 1 \) gives an equivalent to using a double exponential prior, called LASSO regression \[11\]. Still, each \( r \) in (9) can be solved independently, although as penalized regressions. However, the likelihood must be solved iteratively by alternating the maximization over \( \sigma_r \) and the \( \alpha \)'s. See \[8\] for a more detailed description of an algorithm minimizing (9) for both \( \phi = 1 \) and \( \phi = 2 \). However, some argue that doing a simple penalized least squares regression, which is in a sense doing a single iteration of solving (9), is a sufficient approximation \[9\]. The optimal \( \lambda \) in (9) can, as in RDA, be found by minimizing a cross-validation estimate of the classification error.

As in the original LASSO, using \( \phi = 1 \) in (9) might produce a sparsity by forcing many of the \( \alpha \)'s to be exactly zero. This has the benefit of possibly revealing underlying structures in \( T \) and consequently \( \Sigma \) and its inverse.

As can be observed from the estimated correlation matrix in Fig. 1, there are blocks of highly correlated bands. In particular, when the response has a low correlation with the predictors, as in the transitions between these highly correlated blocks, the high intercorrelations between the predictors cause rapid fluctuations in the regression coefficients. An example of this can be seen in the plotted rows of the \( T \) matrix in Fig. 1. The high correlations cause some of the directions in the predictor space to have low variance, which again causes instability in the model estimates. PCR \[10\] regresses the prediction onto the principal components of the predictor covariance matrix, truncating the directions where the predictors have the smallest spread. Thus, both ridge regression and PCR have the common property of shrinking the solutions away from low spread directions.

There are two common approaches to estimate the number of principal components to include in each regression. One is the use of cross-validation on the estimated prediction error, and the other is applying significance tests (F-tests) \[12\]. Using cross-validation estimates of the classification error, as is done in obtaining \( \lambda \) in (9), to determine the number of principal components is difficult because of the interdependence of the number of components in each regression. That is, the number of components must clearly be allowed to differ between each regression, but each regression cannot be independently checked for optimal classification rate. However, the \( p \)-values in the significance tests can be based on cross-validation estimates of the classification error, although at the cost of severely increasing the computation times.
B. Selection of Predictor Subsets

Another approach to low sample regression and problems of predictor collinearity is feature selection. Because our goal is to find sparse models, the focus has been on the SFS. Subset selection has the advantage that it might produce interpretable structures in the matrix, in addition, to increase the generalized prediction accuracy.

The subset selections are applied on each regression, i.e., each row of \( T \), independently. As in the case of PCR, the common techniques for finding which elements to include are cross-validation on the prediction accuracy and the use of F-tests.

1) Restrictions and A priori Knowledge in Structured Regressions: In [18], we proposed using the heuristic of selecting and estimating subdiagonals of \( T \) while keeping the rest zero. In time-series terminology, this corresponds to including only certain lags in the regressions. We refer to this method as sparse Cholesky triangles for inverse covariance estimates (STIC). The approach was evaluated on full data sets with quadratic Gaussian likelihood classifiers and is, in this paper, placed in a linear classifier setting using (very) few training samples.

Several authors suggest that the strong correlations between neighboring features in hyperspectral data indicate that long-range correlations could be ignored. Jia and Richards [19] use this assumption to factorize a full-dimensional density estimate by assuming that the covariance matrix is block diagonal. In [5] and [20], the assumption is the rationale behind the application of low-order antedependence models, i.e., forcing all but the first few subdiagonals of \( T \) to be zero, which again can be seen as using low-order generalized autoregressive (AR) models. In practice, the implementation of such antedependence models can be achieved by limiting the search to contiguous subdiagonals in STIC. A moderately more flexible model can be achieved by letting the order of the AR model differ between each regression, i.e., between each row of \( T \). The latter approach can be obtained by placing simple restraints on the SFS.

Including features which are too correlated might reduce generalization performance in classifiers [21]. In the regression setting of this paper, we can easily apply simple a priori element selection rules limiting this correlation effect. One such heuristic could be that the first subdiagonal of \( T \) (the regression coefficients of neighboring features) should be forced to zero and that two neighboring elements of \( T \) should not be simultaneously selected. This would limit the effect of the highly correlated neighboring features and possibly increase generalization performance.

It is natural to ask if the heuristics and approaches for feature selection really select the most informative elements for classification. The problem is twofold; we lack the knowledge of how many parameters are correct in the sense of being informative in the classification, and the element selection itself might be spurious particularly in the cases of small sample sizes [22]. The latter effect on classification performance can, to some extent, be measured by distributing the selected elements in each row randomly and studying the impact of this on the classification rate.

The approaches to regularizing and structuring that are discussed in this section can be found listed in the left-hand column of Table I.

V. Experiments

A. Data Sets

To give a thorough evaluation of the discussed regression approaches, we have performed experiments on four hyperspectral images captured with different sensors and of various sceneries. The number of spectral bands range from 71 to 171. The first image, Pavia [23], is of an urban scene in Italy, captured by an airborne sensor under the HySens project on June 8, 2002. This data set consisted originally of 80 bands, but the last eight bands, which capture thermal infrared, were excluded from this study together with an extremely noisy band at 1.9580 \( \mu \text{m} \). Seventy-one bands in the range of 496–2412 \( \mu \text{m} \) are used. The image has a pixel size of 2.6 m, and the data are reflectance estimates. The ground truth is divided into nine classes. The second image, Rosis [24], was captured by an airborne sensor (ROSIS-02) during the European Multisensors Airborne Campaign (EMAC-94) on May 10, 1994. The location is the Fontainebleau forest, south of Paris, containing ground-truthed areas of oak, beech, and pine trees. The data set has 81 spectral bands ranging from 430 to 830 nm. The image has a pixel size of 5.6 m and contains reflectance estimates without atmospheric correction. The third and fourth images, Botswana and Kennedy Space Center (KSC), are intended for vegetation inventory [25]. The former was captured by the Hyperion sensor aboard the NASA EO-1 satellite over the Okavango Delta, Botswana, on May 31, 2001, has a pixel size of 30 m, and the labeled data consists of 14 classes. This data set was preprocessed by the University of Texas Center for Space Research (CSR), where uncalibrated and noisy bands that covered water absorption were removed. The number of raw radiance bands used is 145 and covers the range of 448–2355 nm. The latter data set was captured by an airborne sensor over KSC at Cape Canaveral, FL, on March 23, 1996, has a pixel size of about 20 m, and the ground truth consists of 13 classes. Preprocessing of the data was performed by CSR, including atmospheric correction and reflectance estimation in addition to removal of the effects of bad detectors, interdetector miscalibration, and intermittent anomalies. The 171 bands are in the spectral range of 409–2438 nm. All the above data sets are well known, and the listed references are publications where these data sets are used with various classification algorithms. Red–green–blue composites of two of the used images are shown in Fig. 2.

Each data set was divided into two equally sized, spatially separate, and mostly disjoint training and test sets. From the half corresponding to training data, five repeated experiments were created by random sampling. The average performances on test data over all classes and all experiments are reported. The number of elements used for training the classifiers is, in this paper, related to the dimensionality of the corresponding training set. Results from training set sizes, covering all the classes, ranging from 0.5 to 8 times the dimensionality of the sets are studied.

All data sets were normalized by subtracting the total mean and rescaling the mean within-class variances to one. This transformation has no effect on the unbiased ML classifiers, but the data must be standardized for the penalized and subset
regression approaches. In the case of the SVM-RBF, the data was, as per standard operating procedures [26], normalized to a domain of (0, 1) to avoid numerical problems when evaluating the inner products.

B. Experimental Methodology

The $\lambda$ value in the ridge and LASSO method of (9) was estimated by the use of tenfold cross-validation on the classification error. In the case of selecting the best predictor subset in each regression independently, the fast and simple SFS method has been applied. In the subset selection approach and the PCR, the number of elements included and the number of principal components, respectively, were guided by statistical F-tests. The significance level (p-value) in the F-tests was chosen to be 0.025 for the SFS and 0.10 in the PCR approach. The PCR does not have to do multiple comparisons and could therefore have a larger significance level. Using the number of elements selected by SFS as an ad hoc estimate of the optimal number of predictors in each row, we also ran experiments picking the same number of random elements for each row. The importance of the exact choice of predictors is evaluated by the latter experiment.

VI. Results

As can be seen in Fig. 3, most of the regression approaches have a lower classification error than the original LDA, and as the sample count decreases, the differences increase. When the sample count is high, the PCR exceeds LDA in classification error. The LDA most often fails completely when the sample count is less than two times the dimension. Some variability can be seen between the ranking of the different regression approaches, but ridge regression is generally among the best performers, with LASSO a close second. In these low-sample cases, the SVM classifier is often outperformed by all except LDA.

An illustration of how the ridge approach helps generalize the covariance estimates and thus produce more useful discriminating variates is found in Fig. 4. Here, the estimated covariance matrix has been used in calculating the linear transformation giving the Fisher’s canonical variates [17], and both training and test data have been projected onto this subspace. Note that regularization apparently improves the generalization performance reflected in better separation between the classes in the test set. The number of training samples used in this example is equal to two times the dimension of the data set ($= 290$).

Fig. 5 shows the selected elements of $T$ using SFS on the Botswana set for the five repeated experiments added together. Studying the nonzero element patterns appearing when using the SFS and LASSO reveals that the number of nonzero elements increases with increased data set size as expected, but that little underlying structure is apparent. The structures have, at best, a vague resemblance to the blocklike structures often evident in correlation matrices of hyperspectral images. There is a large spread of element selection between the five experiments, except for a tendency to include near-diagonal elements. The results for the other data sets (not shown) have strong similarities. Even though the nonzero elements selected using SFS and LASSO differ widely between the five experiments, the selected element patterns can be interchanged while (more or less) retaining classifier performance. However, as shown in Fig. 6, spreading the selected elements randomly deteriorates the performance substantially when sufficient amount of training data is available. On the other hand, when the number of data is extremely low, there is an actual increase in performance when randomizing the placement of the selected predictor elements.
Fig. 3. Mean classification error plots for the four different data sets.

Fig. 4. Projections onto the first two of Fisher’s canonical variates for the Botswana set. Notice how the classes in the test data are both more separate and easily discerned in the regularized case.

Fig. 7 shows an example of the resulting classification errors when forcing structure upon the nonzero elements of $T$. In three out of the four data sets in this study, enforcing the strict generalized AR-model structure resulted in higher classification error. The exception was the high-resolution urban landcover data set, Pavia, where the AR-restricted model and its unrestricted counterpart alternated having the lowest error. As the training-set sizes increased, the AR-restrained models even surpassed LDA in classification error. Letting simple a priori knowledge of the impairing effect of too high correlation between neighboring bands guide the SFS, as suggested in Section IV-B1, resulted in improved classification results over the unrestricted SFS.

Table I lists the total execution times for various methods and data sets for running the full experiments. STIC, ridge, and LASSO take considerably longer than RDA, whereas the approaches based on stepwise forward search take less computation time than RDA. It is mainly the use of F-tests, as in our implementations of SFS and PCR, instead of relying on cross-validation on the classification errors to estimate parameters, that gives the runtime reductions. Furthermore, ample savings in execution times are made by using AR-model restraints, thus limiting the search space.

Although the results are not shown in this paper, using leave-one-out cross-validation on the prediction error in controlling the SFS and PCR gave reduced performance compared to using F-tests, both in the case of correct classification rate and execution times.

VII. DISCUSSION AND CONCLUSION

Most of the regression approaches outperformed LDA and SVM, and some did better than RDA when it comes to correct
classification rate. The excellent classification performance of the ridge approach can possibly be attributed to the direct link between the single parameter $\lambda$, which is found using cross-validation on the classification error, and the regularizing effect. Furthermore, it dampens the exaggerated correlations, i.e., the high regression coefficients, stronger than LASSO. Attempts
However, the selected elements cannot be placed totally at random, at least not when there are enough training samples to find some of the most important interband dependencies.

When the sample sizes are extremely low, distributing the selected elements randomly yields better results than doing regular sequential subset selection. An explanation for this is that the few elements selected, chosen for their high correlation with the response variable, are overfitting the training data. By selecting randomly, one reduces this overfitting.

Enforcing strict AR models facilitates interpretation of both $T$ and $\Sigma^{-1}$ but reduces the classification performance, giving rise to a conclusion that the long-range dependencies between bands are needed to build an effective hyperspectral classifier. Including simple a priori guidelines to the SFS approach seems to help reduce the damaging effects of highly correlated neighboring features.


table

<table>
<thead>
<tr>
<th>METHOD</th>
<th>PAVIA</th>
<th>ROSIS</th>
<th>KSC</th>
<th>BOTSWANA</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDA</td>
<td>17</td>
<td>10</td>
<td>91</td>
<td>59</td>
</tr>
<tr>
<td>SFS</td>
<td>3</td>
<td>3</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>SFS (apriori)</td>
<td>3</td>
<td>3</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>SFS (AR)</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>PCR</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>STIC</td>
<td>56</td>
<td>103</td>
<td>1688</td>
<td>971</td>
</tr>
<tr>
<td>STIC (AR)</td>
<td>3</td>
<td>2</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>RIDGE</td>
<td>86</td>
<td>120</td>
<td>2001</td>
<td>903</td>
</tr>
<tr>
<td>LASSO$^a$</td>
<td>861</td>
<td>1300</td>
<td>20590</td>
<td>11350</td>
</tr>
<tr>
<td>SVM$^b$</td>
<td>111</td>
<td>174</td>
<td>686</td>
<td>662</td>
</tr>
</tbody>
</table>

$^a$Tuning of the convergence criteria might severely reduce execution times.

$^b$Note that the grid search was based on the general guidelines presented in [26], and might therefore be broader than necessary for this application.

of using cross-validation to determine the p-values in SFS and PCR, in a way mimicking how $\lambda$ was obtained in ridge, gave only slightly better classification results, at least for the SFS. Thus, the extreme increase in computation time for such an approach does not seem justified.

It is difficult to conclude on the performance of the different approaches, but both ridge and the “guided” SFS may be potential alternatives to RDA. The former slightly improves the classification rate, but at the cost of being computationally intensive, whereas the latter has a moderately worse classification rate, but at the cost of being computationally intensive.

While the nonzero element patterns of $T$ from SFS and LASSO are highly varying, the classification performance stays fairly stable, even when the patterns are interchanged between the five experiments for each training set size in the case of SFS. One explanation for this can be the high level of correlation between the features (bands), hence allowing several configurations capture the same essence of the overall correlations.


Are C. Jensen (S’05) received the M.Sc. degree in computer science from the University of Oslo, Oslo, Norway, in 2003. Since 2004, he has been with the Department of Informatics, University of Oslo, Oslo, where he is currently working toward the Ph.D. degree.

His research interests include pattern recognition, image analysis, and signal processing for imaging.

Ashjorn Berge (S’05–M’07) received the Sivilingeniør (M.Sc.) degree in applied mathematics from Norwegian University of Technology and Science, Trondheim, in 2001 and the Ph.D. degree in image analysis from the University of Oslo, Oslo, Norway, in 2007.

He is currently a Research Scientist with SINTEF ICT, Oslo. His current research interests include advanced pattern recognition and image analysis, mainly focused on statistical classification and object detection and tracking in video.

Anne Schistad Solberg (S’92–M’96) received the M.S. degree in computer science and the Ph.D. degree in image analysis from the University of Oslo, Oslo, Norway, in 1989 and 1995, respectively.

During 1991–1992, she was a Visiting Scholar at the Department of Computer Science, Michigan State University. She is currently an Associate Professor with the Digital Signal Processing and Image Analysis Group, Department of Informatics, University of Oslo. Her research interests include synthetic aperture radar image analysis, oil spill detection, hyperspectral imagery, statistical classification, and data fusion.