Sparse Inverse Covariance Estimates for Hyperspectral Image Classification

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

***Abstract*—Classification of remotely sensed hyperspectral images calls for a classifier that gracefully handles high-dimensional data, where the amount of samples available for training might be very low relative to the dimension. Even when using simple parametric classifiers such as the Gaussian maximum-likelihood rule, the large number of bands leads to copious amounts of parameters to estimate. Most of these parameters are measures of correlations between features. The covariance structure of a multivariate normal population can be simplified by setting elements of the inverse covariance matrix to zero. Well-known results from time series analysis relates the estimation of the inverse covariance matrix to a sequence of regressions by using the Cholesky decomposition. We observe that discriminant analysis can be performed without inverting the covariance matrix. We propose defining a sparsity pattern on the lower triangular matrix resulting from the Cholesky decomposition, and develop a simple search algorithm for choosing this sparsity. The resulting classifier is used on four different hyperspectral images, and compared with conventional approaches such as support vector machines, with encouraging results.**

***Index Terms*—Cholesky decomposition, covariance parametrization, hyperspectral image classification, inverse covariance matrix, sparse regression.**

I. Introduction

Classification of pixels in hyperspectral images is a complex problem. We usually have few samples available for training the classifiers, and the input data are of high dimensionality. Further compounding the problem, features usually exhibit high correlation, adding a redundancy to the data that in some cases may obscure the information important for classification.

When using parametric methods, such as the Gaussian maximum-likelihood (GML) 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. A wealth of approaches for dealing with the curse of dimensionality have been proposed in the literature, ranging from dimensionality reduction of the feature space to regularization of parameter estimates by biasing them toward simpler and more stable estimates.

Direct estimation of the inverse covariance matrix was suggested mainly for computational convenience in [1]. In that paper, it was furthermore noted that for many statistical problems the inverse covariance matrix has many zero or near-zero values, and a direct feature selection approach was applied to choose which elements could be set to zero. Obviously, this approach is computationally infeasible for high-dimensional data with covariance matrices having thousands or tens of thousands of elements. We propose an approach that relies on the fact that a modified Cholesky decomposition of an inverse covariance matrix defines coefficients in a regression [2]. By choosing targets in this regression to be zero, we can find simpler models for the covariance matrix with fewer parameters to estimate. A heuristic is suggested for searching for these parameters, guided by measuring classification performance on a ten-fold cross-validation (10-CV), with the goal of finding sparse inverse covariance matrices where only the elements useful for classification are estimated. By reducing the number of parameters to estimate, variability in these covariance estimates is reduced. Our results suggest that classifiers based on these sparse covariance matrices have improved generalization performance.

The main contribution of this paper is a novel approach for expressing and estimating sparse covariance approximations for high-dimensional classification problems. Our proposed method is an extension of well-known time series theory traditionally used in regression analysis of longitudinal time series. By applying these results in a classification context combined with a few weak assumptions regarding the behavior of hyperspectral data, we can define a novel approach for estimating very simple models for full-dimensional class-conditional covariances. The proposed method is by design intended to be used in full-dimensional space—we seek to avoid the curse of dimensionality by directly reducing the number of parameters to estimate. Traditional dimension reduction approaches seek to indirectly reduce the number of parameters to estimate by a potentially cumbersome feature extraction or selection strategy.

In Section II, we will first present the related work that this paper is based upon. Section III will introduce our model and assumptions. Parameter estimates are discussed in Section IV. In Section V, we study the performance of our method by comparing it with conventional classifiers on four different hyperspectral image classification problems. Section VI discusses our results and proposes extensions.

II. Related Work

In the time series literature, several works focus on using the modified Cholesky decomposition of the inverse covariance matrix $\Sigma^{-1} = LDL^{\mathrm{T}}$ to facilitate modeling of the covariance structures of time series. All these papers are based on a

Manuscript received June 20, 2006; revised October 30, 2006. This work was supported by the Norwegian Research Council.

The authors are with the Department of Informatics, University of Oslo, 0316 Oslo, Norway.

Digital Object Identifier 10.1109/TGRS.2007.892598

well-known result from graphical models for multivariate statistics, see for example [2]. The result states that, by using the Cholesky decomposition, the inverse covariance matrix can be estimated by conditional regressions. These methods are in most cases designed for regression analysis of longitudinal data. Longitudinal data are ordered, a property shared with the spectra measured for each pixel in our images. In all approaches explored in the literature, the focus is on expressing the conditional regressions with penalties on the coefficients, to accentuate some attribute of the covariance matrices, such as the occurrence of zeros. Penalized regressions shrinking parameter estimates toward zero, have been proposed for covariance estimation in regression analysis of time series in [3]–[5]. The resulting covariance estimates from these regressions are results from iterative processes, since penalties are dependent on parameter estimates. Furthermore, in [3]–[5], all parameters in the covariance matrices are estimated, even if many of them are forced to zero. Our proposed approach transfers ideas derived from these papers to the domain of classification, as well as choosing the computationally simpler approach of sparse regressions instead of penalization, thus only estimating the necessary parameters for classification. A similar approach can be found in [6], where a thresholding of the conditional mutual information between all possible elements in the feature vector is used to choose which elements to set to zero. A Bayesian approach is explored in [7], where the covariance is modeled by assuming a prior distribution on each individual element of L_k, modeling the probability of the discrete states nonzero/zero.

For data that has some specific order, for example discretized curves, it is known that strong correlations between neighboring features can be detrimental to classifier generalization. One approach for penalization in parameter estimates is proposed in [8], focusing directly on dampening the effect of strong correlations between neighboring features. Our heuristic for searching for a sparse representation is based on the same intuition; assuming that some of the correlations between features do not give information that helps separating the classes, and thus can be dropped.

Inspired by the observation that neighboring features are usually strongly correlated [9] proposed to ignore the long range correlations in spectra. Assuming that the covariance matrix is block diagonal, they obtain a sparsity in the class-conditional covariance parameter estimates. The choice of blocks is based on a heuristic thresholding of the full-dimensional correlation matrix. Since the inverse of a block diagonal matrix is itself a block diagonal matrix, we can view the model of [9] as a simple approximation to a the proposed model assuming that we only allow banded inverse covariance matrices. However, it is not a full-banded model, because correlations between features in neighboring blocks is ignored. Previous to defining the heuristics for sparsity discussed below, we explored a similar pattern—with less than encouraging results.

Sharing some underlying structure of the inverse covariance matrix, as is proposed in this paper, is similar to applying a regularization to the covariance matrix estimate. Several papers have explored different approaches for stabilizing the estimated class covariance matrices in hyperspectral image classification, based on extension of the well-known regularized discriminant analysis [10]. Examples are leave-one-out covariance (LOOC) estimation, using a mixture of four simplifications of the covariance estimates [11] with extensions [12]. In general, the motivating factor for these methods is that the simpler estimates are less variable, since the ratio of the number of parameters to estimate compared to the number of samples available is significantly lower.

In a previous study [13], we considered the idea of regularization by sharing covariance structure, as an extension of model-based clustering [14]. The structure sharing was obtained by sharing characteristics of the covariance eigendecomposition among all clusters in the data, allowing clusters to share orientation, shape, and volume.

Recently, several papers [15]–[19] have focused on using support vector machine (SVM) classifiers for classification of hyperspectral data. The common claim is that these methods are insensitive to dimensionality issues and overtraining. Other authors argue that in general this is not the case [20], but it is clear that SVM classifiers can be used to design arbitrarily complex decision boundaries. SVM classifiers are related to this paper as both approaches strive to use sparse models to describe the decision boundaries between classes. The SVM classifier uses a tuning parameter acting as a regularizer onto the decision boundary, which needs to be carefully adjusted to ensure good generalization performance. This regularization is measured as a cost of misclassifying a training sample. This parameter is denoted C_{err} in this paper. The SVM is a so-called kernel method, meaning that it measures sample distance in some space implicitly defined by a weighting function. A very 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" to evaluate the distance measure. Thus, there is also another free parameter, the width of the Gaussian kernel γ.

III. Sparse Cholesky Triangles for Class-Conditional Covariance Matrices

Consider a classification problem with k classes, assuming Gaussian class-conditional distributions with mean μ_k and classwise 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. Noting that $-\log|\Sigma_k| = \log|\Sigma_k^{-1}|$, it is clear that there is no need for matrix inversion when classifying data if we have a method for estimating the inverse covariance matrices directly. As we shall see in the following sections, a well-known result in the literature on graphical models [2] can be applied to this end.

A. Parametrization of the Inverse Covariance by the Modified Cholesky Decomposition

Define a parametrization of the inverse covariance matrix by the modified Cholesky decomposition [21]

$$\Sigma^{-1} = LDL^{\mathrm{T}}$$

where L is a lower triangular matrix with ones on the diagonal

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

and D a diagonal matrix. 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 number. Several authors in the time series literature have noted this [3], [6], [7]. 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 x_r based on the $r-1$ preceding features $\{x_1, \ldots, x_{r-1}\}$. In keeping with the earlier notation, and assuming zero mean for readability, this can be expressed as

$$x_r = \sum_{j=1}^{r-1} \alpha_{r,j} x_j + \varepsilon_r \tag{1}$$

where the rth diagonal entry $D_{rr} = 1/\mathrm{var}(\varepsilon_r)$. As long as the diagonal elements of D are positive, any choice of α will still produce a positive definite covariance matrix. Sparsity in the representation of the inverse covariance can thus be obtained by fixing some α to be zero.

B. Search for Sparse Cholesky Triangles for the Classwise Covariance Matrices

As pointed out earlier, [1] proposed to choose the sparsity of the inverse covariance matrices using a sequential forward feature selection. Clearly this is infeasible for high-dimensional data where the number of unique elements in the covariance matrix is in the thousands or tens of thousands; thus we have to resort to a heuristic. The general idea of the proposed method is to find a sufficiently complex covariance matrix to solve our classification problem, by evaluating a search space that is small enough to handle. Several different structures for choosing patterns of zeros were considered when defining this heuristic, among these were block patterns. Most of these gave poor results, and there is no clear intuition describing them. The diagonal pattern described below was found to give a good tradeoff between expected training time and classification performance. Furthermore, there is also a clear intuition behind the diagonal pattern.

Search Algorithm: From the regression formulation in (1) we can argue that $\alpha_{r,j} = 0$ indicates that when predicting x_r, x_j does not carry much interesting information. For all rows, if we were to set the coefficient of a specific preceding feature to zero, we could, using time-series terminology, argue that we ignore a specific lag when predicting the next feature. As an example, if we modeled the covariance matrix according to a first-order autoregressive process, the L would be all zero except the for the elements directly below the diagonal. If we ignore a specific lag for all rows in our sequence of regressions, all elements in an off-diagonal vector in L can be set to zero. The term off-diagonal vector, see Fig. 1, denotes a vector parallel to the diagonal, and corresponds to a specific lag. In Fig. 1, L is sparse, and has only two off-diagonal vectors where we estimate parameters.

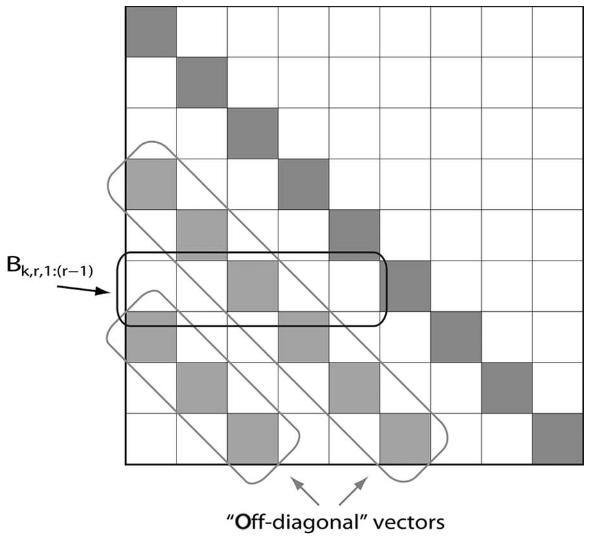

Fig. 1. Illustration of a matrix of correlations L for the inverse covariance matrix. The matrix is lower triangular, with 1 on the diagonal, the elements to estimate is below-diagonal and is represented with B in the text. The sparsity in the covariance estimate is obtained by only estimating the matrix elements in some off-diagonal vectors. The matrix is estimated by a sequence of regressions, one for each row in the matrix B. Thus, for row r, we estimate the elements $B_{k,r,1:(r-1)}$. These regressions can be simplified if we define that all elements not in the chosen off-diagonal vectors are zero.

The general idea is to start by approximating the covariance matrices with the simplest possible model, i.e., diagonal matrices, and add parameters to the approximation until the classification performance of the model no longer improves. With regard to our proposed heuristic, we search for the off-diagonal vectors in the classwise covariance matrices that need to be estimated in order to improve classification performance on the training data. The search, guided by 10-CV as a performance measure, can be described by the following steps.

1) Initialization—Approximate classwise covariance matrices by diagonal matrices, and find 10-CV performance.
2) Search—Select off-diagonal vectors in L_k to be nonzero in a sequential forward manner.
   a) For all zero off diagonal vectors in L_k, evaluate the 10-CV performance gain when allowing each to have nonzero elements.
   b) Add the one off-diagonal vector that gives the largest improvement in 10-CV to the set of off-diagonal vectors to be nonzero in L_k.
   c) Loop from a) until 10-CV performance does not improve further.

This strategy can be viewed as a sequential forward search (SFS) on the off-diagonal vectors. When the optimal number of features in a feature selection is believed to be low, SFS is commonly viewed as a suitable approach. We expect that this applies here as well. Adaptation of more complex strategies for selection of off-diagonal vectors is a subject for future research.

IV. Maximum-Likelihood Inverse Covariance Estimates

As seen in previous sections, the parameters we need to estimate to evaluate the discriminant function is the lower triangular matrix L and the diagonal matrix D. Assuming that classes are Gaussian distributed, closed expressions for these parameters can be found by maximizing the classwise likelihood. By the modified Cholesky decomposition $\Sigma_k^{-1} = L_k D_k L_k^{\mathrm{T}}$, the log-likelihood function for the classwise inverse covariance matrix for class k can be expressed

$$l(\cdot) = \sum_{l=1}^{N_k} \left[-\frac{1}{2} \log\left(|\Sigma_k|\right) - \frac{1}{2} (x_l - \mu_k)^{\mathrm{T}} L_k D_k L_k^{\mathrm{T}} (x_l - \mu_k) \right]$$

where N_k is the number of samples in class k. Express $L_k = I - B_k$, where B_k is a lower triangular matrix with zeros on the diagonal. It is clear that the parameters we need to estimate are the diagonal elements of D_k and the lower triangular elements of B_k. We adopt the following notation: Let $x_{l,r}$ be the rth feature of the lth sample, which gives a vector $x_l = x_{l,1:p}$, where p is the dimensionality of the feature space. Let $B_{k,r,1:(r-1)}$ be the nonzero elements of row r of B_k, i.e., lower triangular elements of the matrix in the given row. See Fig. 1 for an illustration of which matrix elements in B_k that are estimated for row r. Likewise, $x_{l,1:(r-1)}$ is the $r-1$ first features of sample l in the dataset. To simplify the expression, we write $v_{k,l} = x_l - \mu_k$, and $v_{k,l,r}$ and $v_{k,l,1:(r-1)}$ using the same notation as before. We can rewrite the likelihood using these definitions, letting r index the diagonal element $d_{k,r}$ in each row r of D_k. Observe that the log-determinant of Σ_k can be written as the sum of the log of diagonal elements of D_k. ($|L_k| = 1$ by definition.) The log-likelihood, $l(B_k, D_k)$, becomes proportional to

$$\begin{aligned}
&\sum_{l=1}^{N_k} \left[\log\left(|D_k|\right) - \left((I - B_k)^{\mathrm{T}} v_{k,l}\right)^{\mathrm{T}} D_k \left((I - B_k)^{\mathrm{T}} v_{k,l}\right) \right] \\
&= \sum_{l=1}^{N_k} \left[\sum_{r=1}^{p} \log d_{k,r} \right. \\
&\qquad \left. - \sum_{r=1}^{p} \left[\left(I - B_{k,r,1:(r-1)}^{\mathrm{T}} \right) v_{k,l,r} \right]^2 d_{k,r} \right] \\
&= \sum_{l=1}^{N_k} \left[\sum_{r=1}^{p} \log d_{k,r} \right. \\
&\qquad \left. - \sum_{r=1}^{p} \left(v_{k,l,r} - B_{k,r,1:(r-1)}^{\mathrm{T}} v_{k,l,1:(r-1)} \right)^2 d_{k,r} \right].
\end{aligned}$$

The maximum-likelihood estimate for the diagonal elements of D_k can be found by differentiating the log-likelihood with reference to each diagonal element $d_{k,r}$. This gives the following estimates for $r = \{1, \ldots, p\}$:

$$d_{k,r} = \frac{N_k}{\sum_{l=1}^{N_k} \left[v_{k,l,r} - B_{k,r,1:(r-1)}^{\mathrm{T}} v_{k,l,1:(r-1)} \right]^2}.$$

Furthermore, we find the estimate of B_k rowwise by differentiating the log-likelihood by $B_{k,r,1:(r-1)}$ and set to zero. This gives

$$\sum_{l=1}^{N_k} \left[d_{k,r} \left(v_{k,l,r} - B_{k,r,1:(r-1)}^{\mathrm{T}} v_{k,l,1:(r-1)} \right) v_{k,l,1:(r-1)}^{\mathrm{T}} \right] = 0$$

which after some rearranging leads to

$$B_{k,r,1:(r-1)} = \frac{\sum_{l=1}^{N_k} v_{k,l,r} v_{k,l,1:(r-1)}^{\mathrm{T}}}{\sum_{l=1}^{N_k} v_{k,l,1:(r-1)} v_{k,l,1:(r-1)}^{\mathrm{T}}}$$

which is the result of a regression of $v_{k,l,r}$ onto all previous elements in $v_{k,l}$, i.e., $v_{k,l,1:(r-1)}$.

Sparse Regressions of $B_{k,r,1:(r-1)}$*:* The sequence of regressions can be simplified if we assume that some elements of $B_{k,r,1:(r-1)}$ are always zero. This way we can simply remove the corresponding predictors from $v_{k,l,1:(r-1)}$, and thus only estimate the nonzero parameters. Consider Fig. 1 where it can be seen that for row r of B_k has only one target in the regression defined to be nonzero. The implicit sparsity in the representation of the inverse covariance matrix might give a classifier that is more resilient to low sample counts. The number of samples needed to avoid ill-conditioned matrix inversions in the regressions is likely to be lower than the number of samples needed to make sure the full covariance matrix is nonsingular.

V. Experiments

To give a thorough analysis of the performance of our method, we apply the method on four different hyperspectral datasets with dimension ranging from 71 to 176 bands. Our method, sparse Cholesky triangles for inverse covariance estimates (STIC), is compared with a conventional GML classifier, i.e., quadratic discriminant analysis (QDA) and a SVM classifier using radial basis kernels (SVM-RBF). The motivation for comparing the performance of our suggested method with SVM is mainly since it is a method that is supposedly robust to dimensionality and therefore usually applied to data without previous feature extraction—which is by design the intended use of the proposed method. Furthermore, the apparent popularity of SVM in the literature makes it an important and necessary benchmark classifier for any new method to compete with. As noted in related work, regularization by mixing simpler covariance estimates is a popular approach for attacking the instability in parameter estimates due to dimensionality. These regularization methods can be applied in full dimension, which allows direct comparison with the proposed method. To provide a reference for the performance of STIC compared to such approaches, we report experiments using a covariance approximation called LOOC, [11]. It is argued in [12] that LOOC outperforms the well-known regularized discriminant analysis [10] in most cases.

All datasets were normalized by subtracting the total mean, and rescaling total variances for all features to one. This transformation has no effect on the ML classifiers, since it is just a translation and rescaling of the axes. In the case of the SVM-RBF, the data were, as per standard operating procedures [22], normalized to a domain of $\{0, 1\}$ to avoid numerical problems when evaluating the inner products.

To evaluate the performance of different methods, a standard approach is to separate the available ground-truthed data into roughly equally sized regions for training and testing, and report performance on the test data. As far as possible, we have made sure that all regions for training are spatially disjoint from the regions for testing, to avoid training on neighboring pixels that are correlated with test data. We believe this approach gives a more fair indication of classifier performance, although it does in fact make the results difficult to compare with previous publications on these datasets.

From the regions of the dataset available for training, we designed five repeated experiments by sampling equally sized sets for each class for training in the two cases where there was sufficient ground truth to do this. For each of these five experiments, we tune parameters and choose structure using standard 10-CV on each of the sampled training sets. The average performance and stability reported is based on the test data in these five experiments. For the SVM-RBF classifier, two parameters had to be tuned by grid search, the misclassification cost C_{err}, indicating to the cost function the expense of errors during training, and the width of the Gaussian kernel γ, which attenuates the distance measured between samples. To ensure fair treatment of the SVM-RBF, parameters were chosen by a coarse and then progressively finer grid search using the same cross-validation.

For STIC, the implementation used in the reported experiments uses the same map of which off-diagonal vectors that should be nonzero for all classes (i.e., for all L_k). Thus, even though it may be possible to separate some pairs of classes using less parameters, the end result reported is a compromise, where the number of diagonals chosen is the amount deemed sufficient to solve the more complex classification subproblems.

Due to lack of knowledge about true priors of the ground truth classes in any of the datasets, individual performance on each class and average performance over all classes is reported in the experiments.

In the end of this section, overall similarities of the experiments and algorithm training times are presented and discussed.

A. Forest Type Classification

The first dataset we study was captured by an airborne sensor ROSIS during the European Multisensors Airborne Campaign (EMAC-94) in May 10, 1994. The location is a forest south of Paris, Fontainebleau, containing ground-truthed areas of oak, beech, and pine trees. The dataset has 81 spectral bands with sampling bandwidth from 12 nm in the lower part of the spectrum (430–550 nm) to 4 nm in the upper part (554–830 nm), and a pixel size of 5.6 m. The ground truth is divided into three classes, corresponding to tree type, C_1 oak, C_2 beech, and C_3 pine. The dataset is summarized in Table I. The total available number of ground-truthed samples in the dataset was 16 862. Five experiments were performed where 700 samples for each class was used from the training data. On these 700 samples 10-CV was performed to tune parameters.

TABLE I
GROUND TRUTH DATA FOR THE FONTAINEBLEAU IMAGE (IN NUMBER OF PIXELS). SEVEN HUNDRED SAMPLES PER CLASS WERE USED FOR TRAINING FOR EACH EXPERIMENT

Class	Training	Test
C_1 oak	5195	5518
C_2 beech	2083	2309
C_3 pine	807	950

TABLE II
TEST RESULTS FOR FONTAINEBLEAU IMAGE ON FIVE REPEATED EXPERIMENTS. AVERAGE CORRECT CLASSIFICATION AND ONE STANDARD DEVIATION OF THE EXPERIMENTS IN PERCENT

Method	QDA	STIC	SVM-RBF	LOOC
C_1	76.3 ± 1.0	77.0 ± 1.4	81.2 ± 0.5	80.1 ± 0.5
C_2	75.0 ± 0.3	82.8 ± 0.8	81.9 ± 0.8	77.0 ± 0.3
C_3	92.3 ± 0.7	97.2 ± 0.4	98.0 ± 0.4	96.8 ± 0.6
Average	81.2 ± 0.7	85.7 ± 0.9	87.1 ± 0.6	84.9 ± 0.5

The separation of the forest classes is clearly a complex classification problem, especially between classes C_1 and C_2. For this image around 13 off-diagonal vectors were chosen by cross-validation in each of the five repeated experiments. The number of parameters used in the five models was 2348 on average, which is around 23% of the 10 206 parameters of the QDA model for this dataset. The SVM-RBF parameters were tuned using 10-CV. Averaged over the five experiments the misclassification penalty was found to be $C_{\mathrm{err}} = 2^{7.5}$, and the width of the Gaussians, $\gamma = 0.72$. For all classes, the LOOC covariance estimate was chosen to be a mixture of common covariance and classwise covariance. On average for all experiments with LOOC, C_1 and C_2 used roughly 75% and 60% of the common covariance estimate in the mixture, whereas C_3 used only 25%. As shown in Table II, there is much to be gained by using STIC to build a classifier compared with a QDA classification rule or the LOOC regularization. Still, however, the SVM-RBF classifier is slightly better. C_1 and C_2 exhibits mutual confusion in the confusion matrix. This is most likely due to heavy overlap between the classes, but another factor may be slight non-Gaussianity of these classes.

B. Urban Land Cover Classification

We compare the results from the first dataset with data from an urban scene over Pavia, Italy [23], captured by an airborne sensor DAIS under the HySens project in June 8, 2002. This dataset consists of 80 bands, however the last eight bands are thermal infrared bands, and were excluded from this paper. Furthermore, one band at 1.9580 nm was extremely noisy. The number of bands used is 71.

The ground truth consisted of a total of 14 585 samples describing common urban land cover classes chosen by an analyst. The classes were C_1 water, C_2 trees, C_3 asphalt, C_4 parking lot, C_5 bitumen, C_6 roofs, C_7 meadow, C_8 soil, and C_9 shadow. This dataset is summarized in Table III. Again, to avoid testing on direct neighbors of training data, the ground truth sets were split in spatially separate subregions. For each of the five repeated experiments, a set of 100 samples for each class was chosen from the training set.

Table IV summarizes the results for the Pavia image. The STIC result seems to be more stable and slightly better than

TABLE III
GROUND TRUTH FOR THE PAVIA IMAGE (IN NUMBER OF PIXELS). ONE HUNDRED SAMPLES PER CLASS WERE USED FOR TRAINING FOR EACH EXPERIMENT

Class	Training	Test
C_1 water	202	4083
C_2 trees	205	2302
C_3 asphalt	206	1136
C_4 parking lot	205	1301
C_5 bitumen	204	1630
C_6 roofs	201	132
C_7 meadow	315	2041
C_8 soil	202	491
C_9 shadow	119	159

TABLE IV
TEST RESULTS FOR PAVIA IMAGE ON FIVE REPEATED EXPERIMENTS. AVERAGE CORRECT CLASSIFICATION AND ONE STANDARD DEVIATION OF THE EXPERIMENTS IN PERCENT

Method	QDA	STIC	SVM-RBF	LOOC
C_1	99.7 ± 0.5	99.3 ± 0.5	100 ± 0	96.8 ± 1.5
C_2	94.5 ± 1.7	97.9 ± 0.4	93.3 ± 1.3	95.4 ± 0.8
C_3	85.4 ± 6.0	93.4 ± 1.8	97.2 ± 2.0	97.0 ± 0.7
C_4	91.9 ± 1.7	95.7 ± 2.5	75.2 ± 2.8	94.3 ± 1.7
C_5	85.8 ± 7.3	93.0 ± 2.3	97.1 ± 1.0	98.1 ± 0.5
C_6	64.9 ± 11.8	90.6 ± 5.2	85.2 ± 8.4	71.2 ± 5.0
C_7	99.7 ± 0.2	99.7 ± 0.2	97.6 ± 0.7	99.4 ± 0.4
C_8	86.3 ± 3.9	95.9 ± 1.0	94.0 ± 1.0	97.6 ± 0.7
C_9	50.3 ± 2.1	79.0 ± 2.8	93.1 ± 5.9	67.4 ± 35.0
Average	84.3 ± 3.9	93.8 ± 1.9	92.5 ± 2.5	90.8 ± 5.1

the SVM-RBF result, and clearly better than QDA and LOOC. STIC consistently performs well on all classes except C_9 shadow. This class is arguably a collection of several different land cover types and thus the class distribution might be deviating some from the Gaussian shape assumed in STIC. For the SVM-RBF classifier, parameters were found using cross-validation to be $C_{\text{err}} = 2^{12.3}$ for the misclassification cost and $\gamma = 1.57$ for the width of the Gaussian kernels. For this dataset, all classes except C_1 and C_9 were chosen to be an equal mixture of common covariance and classwise covariance in the LOOC approximations. C_9 used only 30% of the common covariance matrix in the mixtures. C_1 was modeled as 60% classwise covariance mixed with 40% of the diagonal of the classwise covariance. The number of parameters chosen by STIC is on average 4025 over the five experiments. This amounts to 17% of the 23 651 parameters used in a QDA model. This sparsity is also reflected in the number of off-diagonal vectors chosen which is on average 4.8 over the five experiments.

TABLE V
GROUND TRUTH FOR THE BOTSWANA IMAGE (IN NUMBER OF PIXELS)

Class	Training	Test
C_1 water	138	132
C_2 hippo grass	51	50
C_3 floodplain grasses 1	119	132
C_4 floodplain grasses 2	106	109
C_5 reeds	130	139
C_6 riparian	133	136
C_7 firescar	135	124
C_8 island interior	87	116
C_9 acacia woodlands	162	152
C_{10} acacia shrublands	129	119
C_{11} acacia grasslands	141	164
C_{12} short mopane	92	89
C_{13} mixed mopane	139	129
C_{14} exposed soils	45	50

TABLE VI
TEST RESULTS FOR BOTSWANA IMAGE. CORRECT CLASSIFICATION RATES IN PERCENT

Method	QDA	STIC	SVM-RBF	LOOC
C_1	0	100	100	100
C_2	100	100	100	100
C_3	0	100	100	100
C_4	0	97.3	99.1	98.2
C_5	0	95.7	87.8	92.8
C_6	0	96.3	90.4	95.6
C_7	0	99.2	98.4	99.2
C_8	0	100	100	92.2
C_9	0	88.8	91.4	88.2
C_{10}	0	98.3	98.3	98.3
C_{11}	0	93.3	87.2	96.3
C_{12}	0	93.3	94.4	92.1
C_{13}	0	97.7	98.5	94.6
C_{14}	100	100	98.0	96
Average	14.3	97.1	96.0	96.0

C. Vegetation Land Cover Classification

1) Satellite Imagery: A third experiment was performed on a scene acquired by the Hyperion sensor aboard the Earth Observing 1 satellite, of the Okavango Delta, Botswana, in May 31, 2001. The dataset originally consisted of 242 overlapping 10-nm bands covering the visible-near infrared (VNIR) and short-wavelength infrared (SWIR), with 30-m pixels. This dataset was preprocessed by the University of Texas Center for Space Research where uncalibrated and noisy bands that covered water absorption features were removed. The total band count of the dataset is 145. Prior publications on this dataset include [24], following which this dataset was made publicly available. The ground truth consists of a total of 3248 samples describing 14 land cover classes chosen to reflect the impact of flooding on the vegetation in the area [24]. The classes are C_1 water, C_2 hippo grass, C_3 floodplain grasses 1, C_4 floodplain grasses 2, C_5 reeds, C_6 riparian, C_7 firescar, C_8 island interior, C_9 acacia woodlands, C_{10} acacia shrublands, C_{11} acacia grasslands, C_{12} short mopane, C_{13} mixed mopane, and C_{14} exposed soil. This dataset is summarized in Table V. To avoid testing on data with high spatial correlation to the training data, the ground truth set was split into spatially separate subregions. The available training data in some of the classes were extremely low, so there was barely enough data for one experiment. Consequently, Table VI does not include estimates of stability.

The results for the Botswana image are given in Table VI. The classification based on QDA breaks down, since many classes have too few ground truth samples to evaluate the full covariance matrices. As shown in Fig. 2, the breakdown due to sample sparsity can be clearly seen, as the error rate increases when the number of parameters approaches a full QDA model. Considering the curves in Fig. 2, the cross-validation performance seems to be a reasonable estimate of generalization performance. A fairly wide region of the performance curve for the cross-validation indicates good model choices. The simple search algorithm in STIC stops quite early in this region and suggests a very sparse model for the covariance matrix. For this image, eight off-diagonal vectors were chosen by the search,

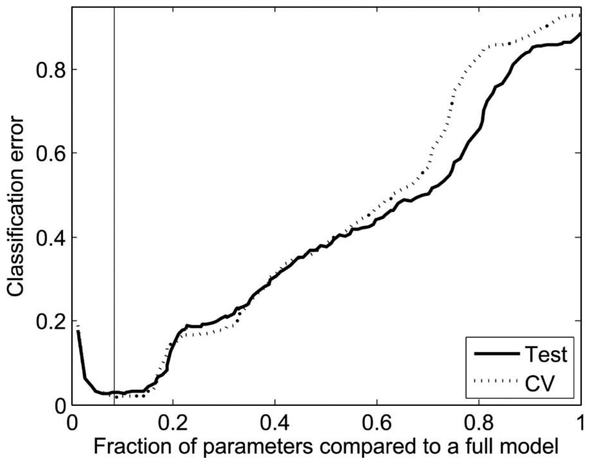

Fig. 2. Average error rates for the proposed method by cross-validation and on test data related to the fraction of parameters of a full model for the Botswana image. The vertical line indicates the point where the search algorithm terminates. Note that the increase in classification error is clearly dependent on the number of parameters estimated.

giving a total of 14 434 parameters to estimate, which is 9.6% of the 150 220 parameters in a QDA model. Parameters for SVM-RBF were found by cross-validation to be $C_{\text{err}} = 2^{11.5}$ and $\gamma = 1.68$. Note that these parameters are quite similar to the ones used on the Pavia image, but that the width of the Gaussian kernel used is roughly doubled compared to the Fontainebleau image. STIC outperforms the SVM-RBF classifier on average, and performs very well in most classes. The poorest performing class, C_9 acacia woodlands is only confused with the class C_6 riparian zone, which might well include pixels covering trees. The LOOC covariance estimate is for all classes based on a mixture of classwise covariance and common covariance estimates, with mixing proportions ranging from 40%–90% of the common covariance, with C_1 water using less (40%) of the common covariance.

2) Airborne Image: Another vegetation land cover classification task is based on a vegetation scene captured by an airborne sensor AVIRIS over Kennedy Space center, Florida, in March 23, 1996. The dataset originally consisted of 224 10-nm bands covering the VNIR and SWIR, with 18-m pixels. Preprocessing and removal of uncalibrated and noisy bands that covered water absorption features was performed by the University of Texas Center for Space Research. The total band-count of the dataset used is 176. The ground truth used consists of a total of 5121 samples describing 13 land cover classes. Prior publications on this dataset include [24], following which this dataset was made publicly available. The classes are C_1 scrub, C_2 willow swamp, C_3 cabbage palm hammock, C_4 cabbage palm/oak hammock, C_5 slash pine, C_6 oak/broadleaf hammock, C_7 hardwood swamp, C_8 graminoid marsh, C_9 spartina marsh, C_{10} cattail marsh, C_{11} salt marsh, C_{12} mud flats, and C_{13} water. This dataset is summarized in Table VII. To avoid testing on data with high spatial correlation to the training data, the ground truth set was split into spatially separate subregions. The available training data in some of the classes were extremely low, so there was barely enough data for one experiment, so Table VIII does not include estimates of stability.

TABLE VII
GROUND TRUTH FOR THE KENNEDY SPACE CENTER IMAGE (IN NUMBER OF PIXELS)

Class	Training	Test
C_1 scrub	379	382
C_2 willow swamp	119	122
C_3 cabbage palm hammock	126	130
C_4 cabbage palm/oak hammock	124	127
C_5 slash pine	80	81
C_6 oak/broadleaf hammock	114	115
C_7 hardwood swamp	52	53
C_8 graminoid marsh	214	217
C_9 spartina marsh	259	261
C_{10} cattail marsh	190	187
C_{11} salt marsh	209	210
C_{12} mud flats	226	236
C_{13} water	456	452

TABLE VIII
TEST RESULTS FOR KENNEDY SPACE CENTER IMAGE. CORRECT CLASSIFICATION RATES IN PERCENT

Method	QDA	STIC	SVM-RBF	LOOC
C_1	0	96.1	100	96.9
C_2	0	95.1	85.2	96.7
C_3	0	93.9	98.6	87.7
C_4	0	81.1	99.5	76.4
C_5	100	76.5	98.5	66.7
C_6	0	72.2	91.8	65.2
C_7	58.5	96.2	91.5	86.8
C_8	0	96.8	70.4	94.5
C_9	0	99.2	78.8	93.5
C_{10}	0	100	94.9	99.5
C_{11}	10.5	98.6	88.7	96.2
C_{12}	0	78.4	75.7	90.3
C_{13}	0	100	94.5	100
Average	13	91.1	89.8	88.5

The classification performance is reported in Table VIII. The high dimensionality of this dataset renders the QDA classifier (using conventional covariance estimates) unusable, since several of the covariance matrix estimates are near singular, resulting in gross overfitting of the training data and consequently poor performance on the test data. However, the STIC approximation finds a useful classifier where we approximate covariance matrices with eight off-diagonal vectors chosen, using in total 8.5% of the parameters of a full QDA model. The number of parameters estimated in STIC is 17 433 compared to 204 776 for QDA. For this dataset parameters for SVM-RBF were found to be $C_{\text{err}} = 2^{15}$ and $\gamma = 11.31$. Note that for this experiment, the width of the Gaussian kernel chosen is quite large compared with the other images, arguably suggesting more linear decision boundaries relative to the other images. The LOOC covariance estimates are on this image a mixture of common covariance and classwise covariance estimates, using 71%–99% common covariance for all classes except C_{13} water. Water C_{13} uses only 40% common covariance. STIC outperforms SVM-RBF and LOOC for this image.

D. Overall Results of the Experiments

Overall, the four experiments indicate that STIC produces strongly performing classifiers while using fairly few parameters. The fraction of parameters chosen for estimation by the search algorithm is, as expected, dependent on the overall

TABLE IX
AVERAGE TIME USED FOR MODEL SEARCH IN THE EXPERIMENTS PERFORMED (IN REAL TIME MINUTES)

Model	STIC	SVM-RBF
Fontainebleau	11.0	276.67
Pavia	4.4	149.02
Botswana	93.9	273.34
KSC	123.7	218.13

complexity of the classification problem. As pointed out earlier, the choice of nonzero off-diagonal vectors is a compromise over all subproblems in the classification task. This seems to be quite clear in the Fontainebleau image, where 23% of the parameters of a QDA model are chosen, in an effort to separate two very similar woodland classes. Another point is that relative to dimensionality, much more data are available for training in the Fontainebleu image.

We note that in most cases, the LOOC regularized covariance estimation tries to stabilize the parameter estimates by using the common covariance estimate. This should not be unexpected, since many classes are very similar, especially in the vegetation classification tasks. The fact that water is a dissimilar class in the two vegetation classification datasets and uses less of the common covariance backs up this assumption. The use of common covariance as a way to stabilize the covariance estimates hints that classification boundaries might be relatively simple for most images.

It is of secondary interest for our experiments, but the SVM-RBF parameters chosen in the four experiments warrant a short comment. In the literature on SVM-RBF, several authors have noted that both the cost of misclassification when training C_{err} and the width of the Gaussian kernel γ influence the complexity of the resulting decision boundary. If we compare the parameters, we observe that very similar parameters are found for the Botswana and Pavia images, whereas the Fontainebleau image suggests a kernel width of roughly half compared to the former images, while the cost of misclassification on the training set is lower. The Kennedy Space Center (KSC) image suggests both high cost of misclassification on the training set and extremely wide Gaussian kernels. Relative to the other images, the Fontainebleau image probably deviates from the normal by needing more complex decision boundaries, however, it is at the same time more regularized than the other images. The parameters of the KSC image, on the other hand, probably reflects fairly linear decision boundaries.

For STIC, the algorithm training time is dependent on input dimensionality. The amount of time needed to estimate classifier parameters when a sparsity pattern is chosen, is negligible. The computational cost comes from the large amount of repeated evaluations in the search algorithm. In Table IX, training time in (real time) minutes is shown, comparing our nonoptimized Matlab implementation of STIC compared with grid searching for SVM-RBF parameters using a Matlab interface to libSVM. Note that training time for STIC increases much faster than the corresponding increase in dimensionality between the images. However, even for the KSC image computation time is still manageable. Training times for SVM are fairly stable, since we evaluate the same grid of parameters for all images.

VI. DISCUSSION

We have proposed a simple algorithm for reducing the complexity of Gaussian ML-based classifiers for hyperspectral data. The main idea is to find a sparse approximation to the inverse covariances of the component distributions used in our classification model. One motivation for developing this approach was to combat the problems with conventional classifiers due to sample sparsity. Our approach is to reduce the number of parameters when the number of available ground truthed samples is low, while sacrificing as little accuracy in modeling the density as possible. The reported experiments show that our method can be expected to perform comparably or better than state of the art conventional classifiers such as SVM, using only a fraction of the full covariance matrices. The performance compared to covariance regularization strategies, in this paper represented by LOOC, seems more than adequate. Our method also performs well in cases where QDA collapses due to sample sparsity. Our method for modeling sparse covariances is also directly applicable to covariance estimates of component distributions in mixture models.

From the search for nonzero diagonals in the Cholesky decomposition, one goal for further research is to develop visualization and exploratory approaches that utilize these choices to give a human analyst hints of the structure of the data and the classification problem, possibly identifying which features or combinations of features, i.e., bands or band-interrelations, that are useful for classification.

Another effect of the proposed method is that it produces probability estimates for each pixel instead of only a decision boundary. After training the classifier, we have approximations of the classwise densities. These probabilities can be used as a pixelwise measure of confidence for class assignment, allowing for evaluation of outliers and doubt cases. Furthermore, the probabilities can be used directly in contextual classifiers.

Ideally, the choice of which correlations that can be ignored should be guided by prior knowledge about the classification problem. One example could be vegetation classes, which probably share the same broad physical characteristics. In this case, it might not be unreasonable to expect that this is reflected in the choice of parameters to estimate. This is a topic for further research. As it is presented, the search algorithm becomes computationally expensive when dimension increases, and the entire space of model choices cannot be searched. A topic for further research is the application of problem specific knowledge to reduce the search space for models.

ACKNOWLEDGMENT

The authors would like to thank the CESBIO institute in Toulouse, France, for providing the Fontainebleau dataset. The Pavia dataset was obtained through Technical Committee 7 of the International Association for Pattern Recognition. The authors would also like to thank the Department of Electronics, University of Pavia, for making this dataset available for research purposes and to Prof. G. Storvik for valuable comments and suggestions.

REFERENCES

[1] A. Dempster, "Covariance selection," *Biometrics*, vol. 28, no. 1, pp. 157–175, Mar. 1972.

[2] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*. Hoboken, NJ: Wiley, 1990.

[3] M. Pouhramadi, *Foundations of Time Series Analysis and Prediction Theory*. Hoboken, NJ: Wiley, 2001.

[4] J. Z. Huang, N. Liu, M. Pouhramadi, and L. Liu, "Covariance selection and estimation via penalised normal likelihood," *Biometrika*, vol. 93, no. 1, pp. 85–98, 2006.

[5] M. Pouhramadi, M. Daniels, and T. Park, "Simultaneous modelling of the Cholesky decomposition of several covariance matrices," *J. Multivar. Anal.*, vol. 98, no. 3, pp. 568–587, Mar. 2006.

[6] J. A. Bilmes, "Factored sparse inverse covariance matrices," in *Proc. IEEE ICASSP*, 2000, vol. 2, pp. 1009–1012.

[7] M. Smith and R. Kohn, "Parsimonius covariance matrix estimation for longitudinal data," *J. Amer. Stat. Assoc.*, vol. 97, no. 460, pp. 1141–1153, Dec. 2002.

[8] T. Hastie, R. Tibshirani, and A. Buja, "Flexible discriminant analysis and mixture models," *J. Amer. Stat. Assoc.*, vol. 89, no. 428, pp. 1255–1270, Dec. 1994.

[9] X. Jia and J. Richards, "Efficient maximum likelihood classification for imaging spectrometer data sets," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 3, pp. 274–281, 1994.

[10] J. Friedman, "Regularized discriminant analysis," *J. Amer. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, Mar. 1989.

[11] J. Hoffbeck and D. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 7, pp. 763–767, Jul. 1996.

[12] B.-C. Kuo and D. Landgrebe, "A covariance estimator for small sample size classification problems and its application to feature extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 4, pp. 814–819, Apr. 2002.

[13] A. Berge and A. S. Solberg, "Structured Gaussian components for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3386–3396, Nov. 2006.

[14] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *J. Amer. Stat. Assoc.*, vol. 97, no. 458, pp. 611–631, Jun. 2002.

[15] J. Gualtieri and R. Cromp, "Support vector machines for hyperspectral remote sensing classification," in *Proc. SPIE 27th AIPR Workshop*, 1998, pp. 221–232.

[16] C. Huang, L. Davis, and J. Townsend, "An assessment of support vector machines for landcover classification," *Int. J. Remote Sens.*, vol. 23, no. 4, pp. 221–232, Feb. 2002.

[17] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[18] G. Camps-Valls, L. Gomez-Chova, J. Calpe, E. Soria, J. Martin, L. Alonso, and J. Moreno, "Robust support vector method for hyperspectral data classification and knowledge discovery," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 1530–1542, Jul. 2004.

[19] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, 2001.

[21] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.

[22] C.-C. Chang and C.-J. Lin. (2001). *LIBSVM: A Library for Support Vector Machines*. [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm

[23] P. Gamba, "A collection of data for urban area characterization," in *Proc. IGARSS*, 2004, pp. 69–72.

[24] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.

Asbjørn Berge (S'05) received the Sivilingeniør (M.Sc.) degree in industrial mathematics from the Norwegian University of Technology and Science, Trondheim, in 2001. Since 2002, he has been working toward the Ph.D. degree at the Department of Informatics, University of Oslo, Oslo, Norway.

His current interests include pattern recognition and image analysis related to processing of hyperspectral remotely sensed data.

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, where he is working toward the Ph.D. degree.

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

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

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