Sparse Covariance Estimates for High Dimensional Classification Using the Cholesky Decomposition

Asbjørn Berge and Anne Schistad Solberg

Department of Informatics
University of Oslo, Norway

Abstract. Results in time series analysis literature state that through the Cholesky decomposition, covariance estimates can be stated as a sequence of regressions. Furthermore, these results imply that the inverse of the covariance matrix can be estimated directly. This leads to a novel approach for approximating covariance matrices in high dimensional classification problems based on the Cholesky decomposition. By assuming that some of the targets in these regressions can be set to zero, simpler estimates for class-wise covariance matrices can be found. By reducing the number of parameters to estimate in the classifier, good generalization performance is obtained. Experiments on three different feature sets from a dataset of images of handwritten numerals show that simplified covariance estimates from the proposed method is competitive with results from conventional classifiers such as support vector machines.

1 Introduction

Many modern pattern recognition problems face the researchers with the problem of feature spaces of high dimensionality coupled with a sparsity of available labeled samples to be used for training. Further compounding the problem in many cases is that features are highly correlated, this 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. Still, many of these methods have slight weaknesses which would be gainful to try to resolve. Among these is the need for inversion of covariance matrices when evaluating the classifier, and estimation of redundant parameters in the full dimensional feature space. Especially when matrices are near-singular, which they tend to be if the ratio between labeled samples

D.-Y. Yeung et al. (Eds.): SSPR&SPR 2006, LNCS 4109, pp. 835–843, 2006.

and dimensionality is ill posed, inversion is plagued with numerical instabilities. Therefore, a direct estimation of the inverse covariance matrix would be useful.

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 with 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. 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, 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. The 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. We propose a heuristic for only estimating the necessary parts of the class-wise covariance matrices based on a simple search algorithm. The reduction in the number of parameters to estimate reduces the variability in the remaining parameters, while we still are using the full dimensional feature space for classification, gaining increased class separability.

2 Sparse Class Conditional Covariance Matrices

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. 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.

2.1 Parametrization of the Inverse Covariance by the Modified Cholesky Decomposition

We decompose the inverse covariance matrix as a modified Cholesky decomposition [2]

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

where L_i 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 row-wise 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], [5]. 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. 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 of $D_{r,r} = \text{var}(\varepsilon_r)$ This parametrization has the effect is that the resulting covariance matrix will still be positive definite, as long as the diagonal elements of D are positive.

2.2 Search for a Sparse Representation of the Class-Wise 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.

Search Algorithm. From the regression formulation in equation 1 we can argue that a zero $\alpha_{r,j}$ indicates that when predicting x_r, x_j does not carry much interesting information. For all rows, if we were to zero the coefficient of a preceding feature, we could, using time-series terminology, argue that we ignore a specific *lag* when predicting the next feature. If we ignore a specific lag for all rows in our sequence of regressions, all elements in an off-diagonal in L can be set to zero. Consider the illustration of a sparse L in figure 1, where the sparse L matrix has only two off-diagonals where we estimate parameters.

We can also observe that zeros in the inverse covariance matrix means conditional independence. It is well known that when there is a zero in position (i, j) of the inverse covariance matrix, x_i and x_j are independent, conditioned on the rest of the features [6]. Informally this means that given all other features, x_i does not carry information regarding x_j and vice versa. For data that has some specific order, for example discretized curves, the intuition that the correlation between neighboring features does not carry information useful for discrimination between classes, and in some cases is even detrimental to classifier results, was established in [7]. Our proposed heuristic for reducing the search space for a sparse representation of the inverse covariance matrix is based on the same intuition, that some of the correlations between features do not help in separating the classes, and thus can be dropped.

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 approximation until the classification performance of the model does no longer improve. With regard to our proposed heuristic, we search for which off-diagonals in the covariance matrices that needs to be estimated in order to improve classification performance on the training data. The search, guided by ten-fold cross-validation (10-CV) as a performance measure, can be described by the following steps:

1. *Initialization* - Estimate diagonal inverse covariances for all classes k, Σ_k^{-1} and calculate 10-CV performance. The parameters to estimate is the variances in D_k.
   No parameters in L_k are estimated.
2. *Search* - Select off-diagonals in L_k to be nonzero in a sequential forward manner
   (a) Find the 10-CV performance gain when adding each individual off-diagonal to the pool of parameters to estimate
   (b) Add the one off-diagonal that gives the largest improvement in 10-CV
   (c) Loop from a) until 10-CV performance does not improve further.

3 Maximum Likelihood Inverse Covariance Estimates

In regard to our motivation in the previous sections, we can develop Maximum Likelihood estimates for the inverse covariance matrix. By the modified Cholesky decomposition $\Sigma_k^{-1} = L_k D_k L_k^T$ [2], the log-likelihood function for the class-wise inverse covariance for class k can be expressed

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

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 r'th feature of the l'th sample $x_l = x_{l,1:p}$, 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 the illustration in figure 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 $v_{l,k} = x_l - \mu_k$, which gives the further expressions $v_{l,k,r}$ and $v_{l,k,1:(r-1)}$ using the same notation as before. We can rewrite the likelihood using these definitions, letting r index diagonal elements $\sigma^2_{k,r}$ of D_k and observing that the log-determinant of Σ_k can be written as the sum of the diagonal elements of D, since the determinant of a matrix product can be written as a product of determinants, and further that $|L_k| = 1$ by definition. The likelihood becomes

$$\begin{aligned} l(\cdot) &= \frac{1}{2}\sum_{l=1}^{N_k} \log(|D_k|) - ((I-B_k)^T v_{l,k})^T D_k((I-B_k)^T v_{l,k}) \\ &= \frac{1}{2}\sum_{l=1}^{N_k}\sum_{r=1}^{p} \log\sigma^2_{k,r} - \sum_{r=1}^{p}[(I - B^T_{k,r,1:(r-1)})v_{l,k}]^2\sigma^2_{k,r} \\ &= \frac{1}{2}\sum_{l=1}^{N_k}\sum_{r=1}^{p} \log\sigma^2_{k,r} - \sum_{r=1}^{p}[v_{l,k,r} - B^T_{k,r,1:(r-1)}v_{l,k,1:(r-1)}]^2\sigma^2_{k,r} \end{aligned}$$

To estimate the elements of the diagonal matrix D_k, differentiate by $\sigma^2_{r,k}$, and set to zero

$$\sigma^2_{r,k} = \frac{N_k}{\sum_{l=1}^{N_k}[v_{l,k,r} - B^T_{k,r,1:(r-1)}v_{l,k,1:(r-1)}]^2}$$

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

$$\sum_{l=1}^{N_k}[\sigma^2_{r,k}(v_{l,k,r} - B^T_{k,r,1:(r-1)}v_{l,k,1:(r-1)})v^T_{l,k,1:(r-1)}] = 0,$$

which after some rearranging leads to

$$B_{k,r,1:(r-1)} = [\sum_{l=1}^{N_k} v_{l,k,1:(r-1)}v^T_{l,k,1:(r-1)}]^{-1}[\sum_{l=1}^{N_k} v_{l,k,r}v^T_{l,k,1:(r-1)}],$$

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

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, $v_{l,k,1:(r-1)}$, and thus only estimate the nonzero parameters. Consider figure 1 where it can be seen that for row r of B_k has only two targets in the regression not defined to be nonzero. The implicit sparsity in the representation of the inverse covariance matrix can be considered a feature selection in each regression. This reduces the size of the matrix to be inverted in the regressions, which might give a classifier that is more resilient to low sample counts, since the number of samples needed to make this matrix inversion ill-conditioned might be much lower than in the conventional case.

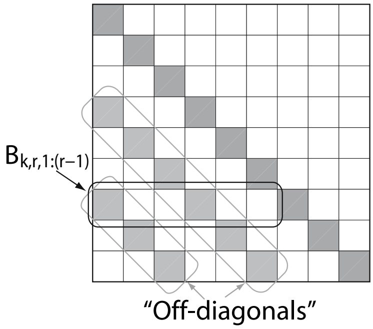
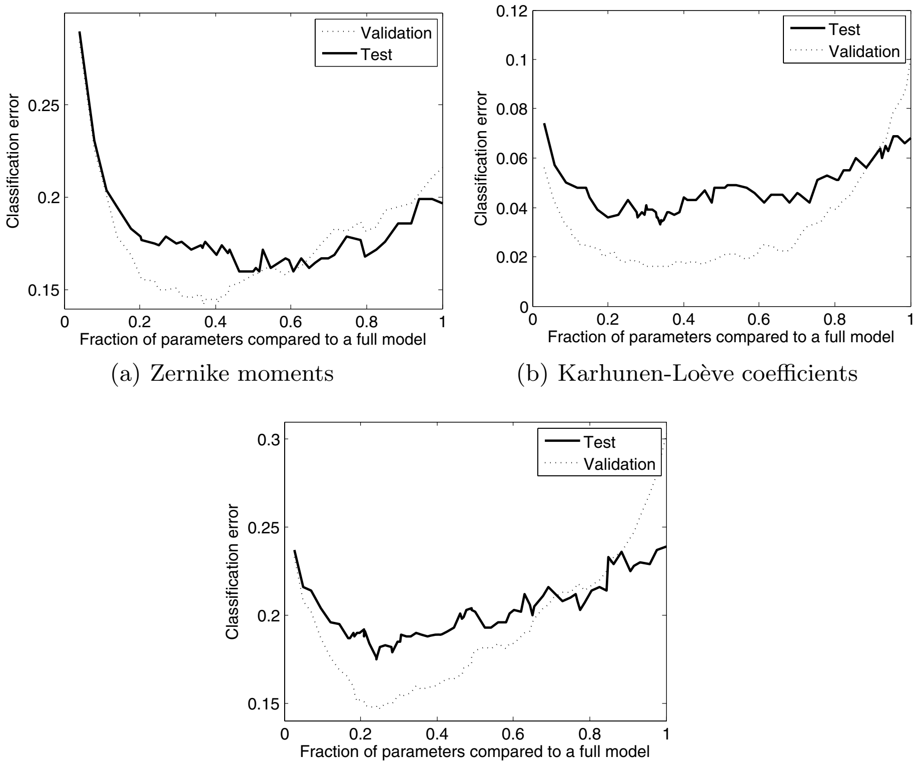


Fig. 1. Illustration of a matrix of correlations, L, for the inverse covariance matrix . The matrix is lower triangular, with 1 on the diagonal, the elements to estimate is below-diagonal and is represented with B in the text. The sparsity in the covariance estimate is obtained by only estimating the matrix elements in *some* off-diagonals. 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-diagonals are zero.

4 Experiments

In our experiments we used the *mfeat* dataset [8], which is a set of images of handwritten numerals. The data consists of 10 classes, each having 200 samples, and the dataset was split randomly in half to generate training and test data. Performance when training was measured using ten-fold cross-validation. From the images, three different feature sets were considered, 47 Zernike moments, 64 Karhunen-Loève coefficients, and 76 Fourier coefficients. The classifier used in our proposed method is a Gaussian Maximum Likelihood classifier, assuming class-wise covariance matrices (GML-quadratic). All covariance matrices are approximated with the same off-diagonals according to the results from the search strategy. In table 1, results of some comparable classifiers on these data are given. The results from the proposed method is included for reference. The classifiers are Gaussian ML classifiers assuming common covariance (GML-linear) and class-wise covariances (GML-quadratic), support vector machines with linear (SVM-linear) and quadratic (SVM-quadratic) kernels, and Parzen density estimation. Note that Zernike moments and Fourier are rotation invariant features, so much of the error in the classification is actually confusion between 6 and 9.

In figures 2(a), 2(b) and 2(c), the error rate by cross-validation and on test data is given as a function of the fraction of covariance elements compared to a full model. A full model estimates the entire covariance matrix for each class, just as a GML-quadratic classifier, but still avoids inverting the covariance matrix. Interestingly, avoiding matrix inversion in the classifier seems to make the classifier slightly more stable than the conventional GML classifier. Figure 2(a) considers the Zernike moment feature set. The classification performance by



(a) Zernike moments

(b) Karhunen-Loève coefficients

(c) Fourier coefficients

Fig. 2. Error rates for the proposed method by cross-validation and on test data compared to the fraction of covariance parameters of a full model for (a) 47 dimensional Zernike moments feature set, (b) 64 dimensional Karhunen-Loève feature set and (c) 76 dimensional Fourier feature set. For the Zernike and PCA feature sets, around 30% of the parameters of a full model seems sufficient for good generalization performance. For the Fourier feature set the number of parameters sufficient for a good classifier is around 25%. These choices is clearly suggested by the cross-validation classification error.

cross-validation has a minimum at 36.8% of the covariance parameters, and the mean result on the test data for that fraction of parameters is 17.4%. The same results for the Karhunen-Loève feature set are given in figure 2(b). The minimum by crosss-validation is here at 29.6% of the parameters, and the mean classification result for this experiment is 3.7%. For the Fourier coefficient feature set the results are shown in figure 2(c). This feature set had a minimum classification error by cross-validation at 25.1% of the parameters, and the classification result on the test set was 18.2%. Note the far right on the figures, since the number of samples available for training is nearing the dimensionality of the dataset, a full covariance model will be near singular and even the proposed model collapses. However, the decline is very graceful, and does not start until 80% of the

Table 1. Error rates (in percent) for classifiers on test data (100 samples per class for training), and on simplified models found by the proposed method on the three feature sets, Zernike moments, Fourier coefficients and Karhunen-Loève coefficients.

Classifier	Zernike	Fourier	K-L
GML-quadratic	19.8	23.9	6.8
GML-linear	18.2	18.5	4.5
SVM-linear	18.3	19.6	6.4
SVM-quadratic	15.7	15.9	2.1
Parzen	18.5	17.0	3.0
Proposed method	17.4	18.2	3.7

parameters of the model is used. These results are summarized in table 1, and compared with results for other classifiers.

All experiments indicated a gradual decline in performance as the number of features estimated increased, however, at the same time the performance curves in figures 2(a), 2(a) and 2(c) indicate that for three different feature sets, there is a fairly wide area where the classification performance on the test set is good. In all the experiments, the minimum classification error by cross-validation occurred in this area.

Considering the results presented in table 1, we observe that the proposed method is certainly competitive with conventional methods.

5 Conclusion and Future Work

Using results from time series analysis, we have proposed a novel approach for estimating sparse covariance matrices in full dimensional feature spaces for Gaussian ML classifiers. Experiments on different feature sets of a handwritten numeral classification problem indicates that it performs equally, or better than conventional classifiers. The results from these initial experiments are encouraging, and suggests the usefulness of further research in this direction.

We envision that this approach will be useful for reducing the number of parameters to estimate in a mixture of Gaussians classifier, where the motivation for using the sparsest possible model for component distributions is even stronger.

Future work with this method will focus on improving the selection heuristic. The present heuristic of choosing off-diagonals in L to estimate is intuitive when there is some clear correlation structure in the data that is present in the entire feature set. An example of this is the strong correlation between neighboring features when using discretized curves as input. Another improvement would be to consider selecting different off-diagonals to estimate for each class.

References

1. Dempster, A.: Covariance selection. Biometrics (1) (1972) 157–175
2. Golub, G.H., Van Loan, C.F.: Matrix computations, 3rd ed. John Hopkins University Press (1996)

3. Smith, M., Kohn, R.: Parsimonius covariance matrix estimation for longitudinal data. Journal of the American Statistical Association **97**(460) (2002) 1141–1153
4. Bilmes, J.A.: Factored sparse inverse covariance matrices. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP'00). Volume 2. (2000) 1009–1012
5. Pouhramadi, M.: Foundations of Time Series Analysis and Prediction Theory. Wiley (2001)
6. Whittaker, J.: Graphical models in applied multivariate statistics. Wiley (1990)
7. Hastie, T., Tibshirani, R., Buja, A.: Flexible discriminant analysis and mixture models. Journal of the American Statistical Association **89**(428) (1994) 1255–1270
8. Jain, A.K., Duin, R.P., Mao, J.: Statistical pattern recognition: A review. IEEE Trans. Pattern Anal. Machine Intell. **22**(1) (2000) 4–37