Discussion: Sparse Discriminant Analysis and Multi Collinearity with Applications to Image Analysis

Line H. Clemmensen[*]

Abstract
In an image analysis context, data often have high correlations due to spatial and spectral relations between pixels. The high correlations are inherited to features when feature extraction is applied to the images. Here, we are concerned with the classification setting, and in particular extending discriminant analysis to the high-dimensional case ($p > n$). We review a variety of sparse discriminant analysis techniques, and their relation to the data of interest. Subsequently, grouping of variables in the elastic net is reviewed as well as the adaptive group elastic net. The paper is a discussion of the methods for high-dimensional multi collinear data sets derived from image analysis problems, where shrinkage is at the core of minimizing the variance and obtaining generalization power as well as interpretable solutions.

Key Words: Classification, High Dimensionality, Image Analysis, Multi Collinearity, Sparse Discriminant Analysis

1. Introduction

Image analysis is increasingly used in industrial as well as research applications, and the problmes are often of a predictive or supervised nature. This paper looks at correlation structures resulting from image analysis problems, and discusses the influence on teh choice of regression and classification algorithms.

This paper examines the correlation structure of data derived from images, and emphasizes the use of shrinkage for such problems where features are highly correlated. A number of sparse dicriminant analysis techniques are reviewed along with an adaptive grouping technique for the elastic net.

2. Image Analysis Examples

This section introduces two examples of predictive scopes using image analysis. The first example derives a smaller set of features from the original images, whereas the second example uses the color intensity in the images as features. The examples serve to demonstrate the strong correlation structures often arising in image anlaysis problems, and thus lay the basis for the further discussions in this paper.

2.1 Spectral images with feature extraction

The motivations for the first example is to predict moisture in sand for concrete production by using imaging techniques such that the predicitons can be performed inline and waiste concrete can be minimized as the proper mixing proportions are estimated correctly. For the purpose, near infrared spectral imaging was used with 9 different wavelengths for fast and cheap acquisition (Clemmensen et al., 2010).

As each image consists of $9 \times 1035 \times 1380 \simeq 9,000,000$ pixels, the dimensionality was further reduced by extracting 1st order statistics like percentiles from each wavelengths

[*] Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads 324/220, DK-2800 Lyngby, Denmark

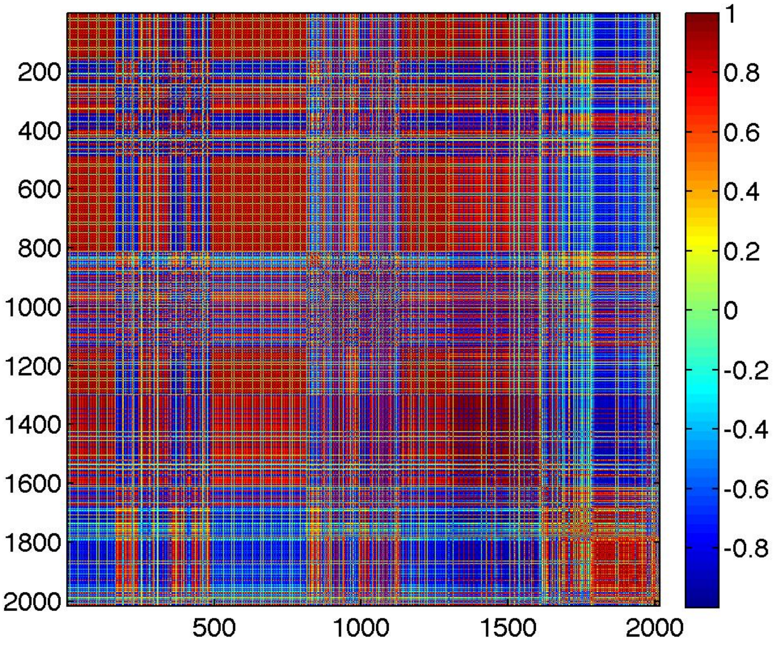

Figure 1: Correlation matrix for sand data.

as well as pairiwse combinations of wavelengths. Finally, texture features were also extracted. The feature extraction amounted to a total number of 2016 features (Clemmensen et al., 2010). The data set has 59 observations Figure 1 illsutrates the correlation structure amongst the extracted features.

Though this data comes from a regression problem, similar feature extraction has proven useful in similar spectral image analysis problems where classification is the task at hand (Clemmensen et al., 2007).

2.2 Standard color images

This example uses standard color images respresented by a red (R), green (G), and a blue (B) color band for each image of three fish species: Cod, haddock, and whiting. The objective is to classify the three species in a fast and objective manner which for example could be implemented in the fish industry or to control fish quota.

A number of image analysis techniques were applied to get a one-to-one correspondance between the fish in the various images, and subsequently the color intensities can directly be used as features (Larsen et al., 2009). As a part of the pixel matching between fish images, the shape of each fish is described using an annotation. To include features of size and shape, these landmark features were used. In total this amounts to 160,072 features, and the data set has 108 observations. Figure 2 illustrates the correlation matrix between the features.

It is clear that there is a strong correlation between R, G, and B features, as patterns are repeated in three blocks, also on the off diagonal. The first few features are shape features and the pattern is distinct for these. Finally, also strong correlations are seen within a color due to the spatial structure of the data, i.e. pixels next to each other have a higher correlation than pixels further apart.

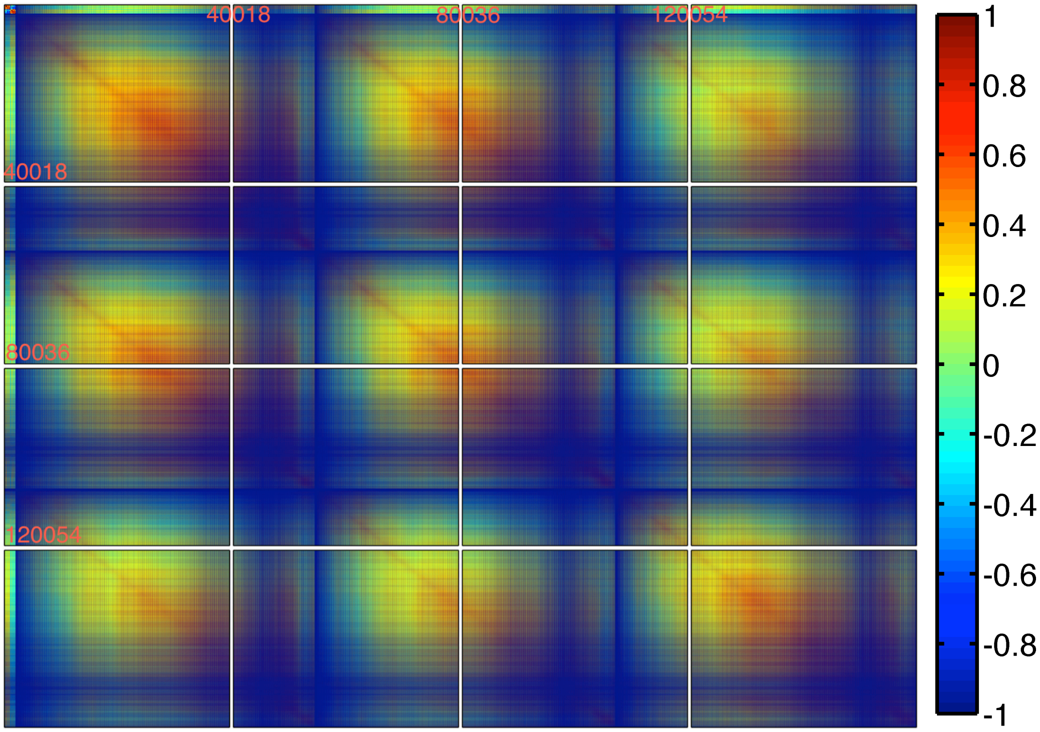

Figure 2: Correlation matrix for fish data.

3. Methods

This section reviews various discriminant analysis techniques for high-dimensional problems, and discusses pros and cons in relation to analysing data from image analysis.

3.1 Discriminant Analysis Techniques

This section reviews some of the recently proposed techniques for performing discriminant analysis in high dimensioanl spaces. The scope is to evaluate their appropriateness in relation to the proposed image analyis problems, i.e. for problems with several classes, more features than observations, and high multi-collinearity.

3.1.1 Independence

Nearest shrunken centroids (NSC, Tibshirani et al. (2003)) and Penalized linear discriminant analysis (PLDA, Witten and Tibshirani, 2011, Witten and Tibshirani (2011)) uses the diagonal of the within-class covariance matrix, and thus assumes independence between the features. NSC uses soft thresholding and PLDA uses an ℓ_1-norm (lasso, Tibshirani (1996)) to obtain sparsity in the feature space.

3.1.2 Soft thresholding

Regularized discriminant analysis (RDA, Guo et al. (2007)) estimates the within-class covariance matrix as a weighted average of the full estimate and the diagonal estimate, and introduces sparsity through soft thresholding of the covariance projected means ($\Sigma^{-1}\mu$). Soft thresholding is comparable to a lasso shrinkage. With a regularized estimate going from the full covariance matrix to the diagonal, the method is able to solve a wide variety of problems.

3.1.3 Elastic net regularization

Sparse discriminant analysis (SDA, Clemmensen et al. (2011)) reframes the classification problem into a regression problem using optimal scoring, and then uses the ealstic net (Zou and Hastie, 2005) to include shrinkage and sparisty to the solution. Shrinkage is highly recommended to include when the features are correlated, as is the case in many image analysis problems.

3.1.4 Thresholding

Sparse linear discriminant analysis by thresholding (SLDAT, Shao et al. (2011)). This method uses thresholding of both the wihtin-class covariance matrix and the class means. They showed that under certain conditions, their method is asymptotically optimal. This method also solves a wide variety of problems going from a full estimate of the covaraince matrix to a diagonal estimate. However, the thresholding does not introduce additional shrinkage, and often thresholding methods are known to perform better if re-estimated after thresholding due to multi-collinearity.

3.1.5 Additional ℓ_1-penalties

ℓ_1-Fishers discriminant analysis (FSFD, Wu et al. (2008)) introduces a lasso penalty to the Fisher's formulation of discriminant analysis for two-class problems. Similar to PLDA, though without a diagonlization of the within-class covarince matrix.

The direct sparse discriminant analysis (DSDA, Mai et al. (2012)) is a method for binary classification with a penalization on the least squares formulation of two-class linear discriminant analysis. In the paper, a lasso penalty was chosen, thus only mild shrinkage is performed, but an elastic net penalty could also be used. DSDA achieves the Bayes error rate as n goes to infinity. Recently, Mai and Zou (2014) proved that for two-class problems, the three methods SDA, FSDA, and DSDA have equivalent normalized solutions.

The ℓ_1-norm penalty works for $p > n$ cases, but the grouping properties of the ℓ_2-norm are desirable when high correlations exist.

3.1.6 Further reading

A more thorough discusssion on some of the discriminant analysis techniques including numerical simulation results can be found in Clemmensen (2013). For the more theoretical comparison see Mai and Zou (2014).

3.2 Grouping in The Elastic Net

Several of the methods above, for example SDA and DSDA can or could use the elastic net penalty to reduce the dimensionality. The elastic net consists of both an ℓ_2-norm shrinkage (Hoerl and Kennard, 1970) and an ℓ_1-norm feature selection penalty (Tibshirani, 1996) of the parameters. The ℓ_2-norm shrinkage has a natural grouping of features that are highly correlated, such that they will each contribute in a weighted form to the solution (Zou and Hastie, 2005; Hastie et al., 2009).

Several explicit grouping methods have been proposed, see to Clemmensen (2014) for more details. To my knowledge none of them are designed for adaptively grouping the variables, i.e. when the groups are not known a priori, or for truely high-dimensional problems. For this reason the adaptive group elastic net considers a simple definition of numerically equal correlations as

$$if \quad c_j - \delta \leq c_i \leq c_j + \delta \quad \Rightarrow \quad c_j \equiv c_i \quad , \tag{1}$$

i.e. we accept that c_j equals c_i, and adds groups of such features rather than a single feature at a time.

For a given regression problem with data matrix $\mathbf{X}$, response $\mathbf{y}$, the adaptive grouping algorithm as defined in Clemmensen (2014) is

Algorithm 1 Group elastic net algorithm

```
1. Require $\mathbf{X}_{p\times n}$, $\mathbf{y}_{n\times 1}$, and $\delta$.
2. Ensure $\sum_i y_i = 0$, $\sum_i x_{ij} = 0$, $\sum_i x_{ij}^2 = 1, \forall j$.
3. Initialize the active set $\mathcal{A} = \emptyset$, the inactive set $\mathcal{I} = \{1, ..., p\}$, the correlation threshold $r_t = 1 - \delta$, the iteration number $k = 0$, and the current estimate $\mu_k = \mathbf{0}$.
4. While early stopping criterion not met
    (a) Compute the current correlations $c_j = \mathbf{x}_j^T(\mathbf{y} - \mu_k)$, $j \in \mathcal{I}$, with maximum corerlation $C = \max(\text{abs}(c_j))$ and corresponding index $\mathcal{M} = \arg\max_j(\text{abs}(c_j))$.
    (b) Identify the set of variables with correlations of equivalent size to $C$, as $\mathcal{P} = \text{find}(\text{abs}(c_j - C) \leq \delta)$ , $j \in \mathcal{I} \setminus \mathcal{M}$.
    (c) Compute the correlations between the variable with maximum current correaltion and all variables in the identified set $\mathcal{P}$, $\mathbf{r} = \mathbf{X}_{\mathcal{M}}^T \mathbf{X}_{\mathcal{P}}$. Find the the set of grouped covariates as the variables which have suitable correlation sizes $\mathcal{J} = find(\mathbf{r} > r_t)$.
    (d) Compute the step length $\gamma$ and the equiangular direction $\mathbf{u}_{\mathcal{A}}$ using 2 and 3, and update the current prediction $\mu_{k+1} = \mu_k + \gamma \mathbf{u}_{\mathcal{A}}$, and the set of active variables $\mathcal{A}_{k+1} = \mathcal{A}_k \cup \mathcal{M} \cup \mathcal{J}$.
    (e) End of iteration, $k = k + 1$.
5. Return the active set of variables and the predictions.
```

This algorithm builds on the least angle regression and selection (LARS, Efron et al. (2004)) algorithm. LARS includes the variable(s) with an equal correlation and proceeds in the direction of the equiangular vector $\mathbf{u}_{\mathcal{A}}$ of the active set $\mathcal{A}$, given by:

$$\mathbf{u}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}} \mathbf{w}_{\mathcal{A}}, \tag{2}$$

where the covariates of the active set of variables is denoted as $\mathbf{X}_{\mathcal{A}}$, and $\mathbf{w}_{\mathcal{A}} = A_{\mathcal{A}} \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}} \mathbf{1}_{\mathcal{A}}$, with $A_{\mathcal{A}} = (\mathbf{1}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}} \mathbf{1}_{\mathcal{A}})^{-1/2}$, and $\mathbf{1}_{\mathcal{A}}$ is a vector of ones with the size of the active set $\mathcal{A}$. The length of each step is given by:

$$\gamma = \min_{j \in \mathcal{I}} \left\{ \frac{C - c_j}{A_{\mathcal{A}} - (\mathbf{X}_{\mathcal{I}}^T \mathbf{u}_{\mathcal{A}})_j}, \frac{C + c_j}{A_{\mathcal{A}} + (\mathbf{X}_{\mathcal{I}}^T \mathbf{u}_{\mathcal{A}})_j} \right\}, \tag{3}$$

where $\mathcal{I} = \mathcal{A}^c$ is the set of inactive variables and complementary to the set of active variables, and thus $\mathbf{X}_{\mathcal{I}}$ denotes the covariates for the inactive set of variables.

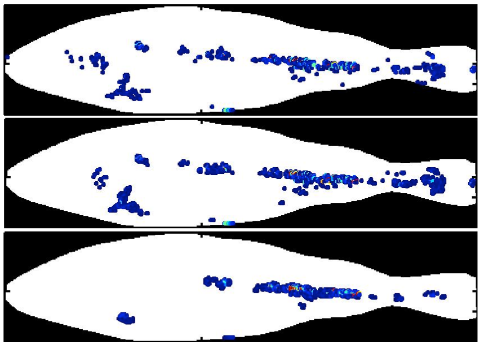

Figure 3: Selected blue features for the 1st discriminant direction for 100 bootstrap repetitions without grouping (top), with mild grouping (middle), and strong grouping (bottom). For both algorithms, early stopping was set to 60 features and the shrinkage parameter was set to 10^6, and for the grouping $\delta = 0.05$ and $\delta = 0.1$.

4. Numerical Results

Various simulation comparisons as well as comparisons on the given examples have been reported in Clemmensen et al. (2010), Clemmensen et al. (2011), Clemmensen (2013) and Clemmensen (2014) which illustrated that shrinkage was a good idea, and that it was important to estimate a full though shrunken correlation structure in the discriminant analysis. We will here revisit the fish image example to give some additional numerical results and interpretations.

4.1 Fish data

The grouping algorithm is illustrated for the fish data in Figure 3. The test classification errors were comparable: for the mild grouping it was around 4.4% with a standard deviation of 3.3% whereas without grouping it was around 4.2% with a standard deviation of 3.3%. The strong grouping had a misclassification rate of 5.8% with a standard deviation of 3.8%.

The dark line down the mid of two of the fish species is a clear interpretation of the discriminant power of the 1st discriminating direction, and the picture gets clearer as the grouping is increased.

5. Recommendation

The author recommends using shrinkage in algorithms for regression or classification in high-dimensional image analysis prolems where strong multi collinearity exist. Its grouping and shrinkage abilities can aid in both generalization and interpretation.

References

Clemmensen, L., Hansen, M., Ersbøll, B., Frisvad, J., Jan 2007. A method for comparison of growth media in objective identification of penicillium based on multi-spectral imaging. Journal of Microbiological Methods 69, 249–255.

Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B., 2011. Sparse discriminant analysis. Technometrics 53 (4), 406–413.

Clemmensen, L. H., 2013. On discriminant analysis techniques and correlation structures in high dimensions. Tech. Rep. 2013-04, Department of Applied Mathematics and Computer Science, Technical University of Denmark.

Clemmensen, L. H., 2014. Selecting groups of covariates in the elastic net. Tech. Rep. 2014-15, Department of Applied Mathematics and Computer Science, Technical University of Denmark.

Clemmensen, L. H., Hansen, M. E., Ersbøll, B. K., 2010. A comparison of dimension reduction methods with applications to mutli-spectral images of sand used in concrete. Machine Vision and Applications 21 (6), 959–968.

Efron, B., Hastie, T., Johnstore, I., Tibshirani, R., 2004. Least angle regression. Annals of Statistics 32, 407–499.

Guo, Y., Hastie, T., Tibshirani, R., 2007. Regularized linear discriminant analysis and its applications in microarrays. Biostatistics 8 (1), 86–100.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, 2nd Edition. Springer.

Hoerl, A. E., Kennard, R. W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Larsen, R., Olafsdottir, H., Ersbøll, B., 2009. Shape and texture based classification of fish species. In: 16th Scandinavian conference on image analysis. Springer Lecture Notes in Computer Science.

Mai, Q., Zou, H., 2014. A note on the connection and equivalence of three sparse linear discriminant analysis methods. Technometrics 55 (2), 243–246.

Mai, Q., Zou, H., Yuan, M., 2012. A direct approach to sparse discriminant analysis in ultra-high dimensions. Biometrika 99 (1), 29–42.

Shao, J., Wang, G., Deng, X., Wang, S., 2011. Sparse linear discriminant analysis by thresholding for high dimensional data. The Annals of Statistics 39 (2), 1241–1265.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of Royal Statistical Society - Series B 58 (No. 1), 267–288.

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2003. Class prediction by nearest shrunken centroids, with applications to dna microarrays. Statistical Science 18, 104–11.

Witten, D., Tibshirani, R., 2011. Penalized classification using fisher's linear discriminant. Journal of the Royal Statistical Society, Series B.

Wu, M., Zhang, L., Wang, Z., Christiani, D., Lin, X., 2008. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set pathway and gene selection. Bioinformatics 25, 1145–1151.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of Royal Statistical Society - Series B 67 (Part 2), 301–320.