Structures of Covariance Matrix in Handwritten Character Recognition

Šarūnas Raudys and Masakazu Iwamura

Vilnius Gediminas Technical University, Saulėtekio 11, Vilnius, Lithuania
Tohoku University, Aoba 05, Aramaki, Aoba-ku, Sendai, 980-8579 Japan
E-mail: raudys@ktl.mii.lt, masa@aso.ecei.tohoku.ac.jp

Abstract. The integrated approach is a classifier established on statistical estimator and artificial neural network. This consists of preliminary data whitening transformation which provides good starting weight vector, and fast training of single layer perceptron (SLP). If sample size is extremely small in comparison with dimensionality, this approach could be ineffective. In the present paper, we consider joint utilization of structures and conventional regularization techniques of sample covariance matrices in order to improve recognition performance in very difficult case where dimensionality and sample size do not differ essentially. The techniques considered reduce a number of parameters estimated from training set. We applied our methodology to handwritten Japanese character recognition and found that combination of the integrated approach, conventional regularization and various structurization methods of covariance matrix outperform other methods including optimized Regularized Discriminant Analysis (RDA).

1 Introduction

One of characteristic elements of modern pattern classification tasks is extremely large number of features that are of the similar origin. An example is classification of handwritten Japanese characters. Since the features are mutually correlated, one cannot ignore the correlations for designing the pattern classification algorithm. To reduce c*omplexity/sample size problems*, one needs to structurize covariance matrix (CM), i.e. describe it by small number of parameters. Two decades ago such approach has been used for classification of time series [1, 2], 2D remote sensing image classification [3-7]. Structurization approach has been utilized also in recognition of handwritten Japanese characters, too.

In many real world problems, distributions density functions of single features have clear deviation from Gaussian law. Promising way to solve such pattern recognition tasks is utilization of artificial neural networks based methods which do not require assumptions about type of distribution density functions of input features. In case of successful training, often one obtains good results. There are two main difficulties to apply such methods. First, results obtained depend on initial conditions (weight vector). Secondly, if input features are highly correlated,

in high-dimensional situations the data becomes almost singular. This makes training become very slow.

A way to diminish the perceptron initialization problem and singularity of the data is the integrated approach of statistical and neural networks based methods [8-10]. Instead of using statistical estimate of CM to design the statistical classifier (denoted by CL_S), we use CM for data whitening transformation. In subsequent training of SLP, this strategy leads classifier CL_s just after the first bach-mode training with zero valued initial weight in the transformed feature space.

If the assumption of structure of the CM is truth and *sample size/complexity relationship* is sufficiently high, we have a good start to train the perceptron further. Good initialization leads to high-quality result if one stops training in a right moment [11]. Moreover, data whitening speeds up training process.

The integrated approach has been derived with the assumption that CM's of both classes are the same. This approach could be ineffective because of unequal CM's. It also could be ineffective when the assumptions of the structures of the CM are far from reality, due to use of wrong covariance structures or use of unreliable estimates which are calculated from small samples for the dimensionality. To improve effectiveness of the integrated approach, one can introduce additional regularization of the CM. An objective of the present paper is to investigate joint application of the CM regularization, standard and special CM structures designed for 2D spatial image recognition to the integrated approach for discrimination of handwritten characters. We performed experiments with similar pairs of Japanese characters (Fig. 1), however, our methodology is not application specific.

鳥 烏	采 釆	伸 仲	乎 平	東 朿	熊 態	帥 師
栗 粟	ぽ ぼ	麿 磨	棒 捧	問 間	椎 推	肋 助

Fig. 1. Fourteen pairs of similar Japanese characters.

The standard CM structures are widely used structures, and the special ones are prepared with taking into account nature of feature vectors of 2D images: distant pixels in the 2D image have less important correlations. The covariance matrices of similar classes are expected to be similar as well as the postulated correlations structures be truthful. We performed our investigation of 2 class discrimination in very difficult condition where the number of sum of sample sizes, $n=N_1+N_2$, and dimensionality, p, are approximately equal. N_i is training sample size of class i.

2 Integrated Approach of Statistical Estimators and Artificial Neural Networks

2.1 Standard Fisher Linear Discriminant Function

The standard Fisher linear discriminant function is the most important rule to classify two categories, and offered the opportunity to give birth to the integrated

approach. Suppose both pattern classes share a common covariance matrix. Denote the pooled sample covariance matrix by **S** and the sample mean vectors of two classes by $\bar{\boldsymbol{x}}^{(1)}$ and $\bar{\boldsymbol{x}}^{(2)}$. Then, allocation of a p-variate vector $\boldsymbol{x} = (x_1, \cdots, x_p)^T$ is performed according to a sign of discriminant function (DF)

$$g(\boldsymbol{x}) = \left(\boldsymbol{x} - \frac{1}{2}\left(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)} \right) \right)^T \mathbf{S}^{-1} \left(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)} \right). \quad (\mathbf{1})$$

Instead of **S**, a "better" (simplified) estimate of the covariance matrix (say $\mathbf{S}_\mathrm{S}$) could convert DF (1) into (another) statistical classifier CL_S with possibly enhanced small sample properties.

2.2 Integrated Approach

In the integrated approach, the learning process consists of two stages: data whitening transformation by statistically estimated CM, and subsequent learning of SLP. Recognition is performed with trained SLP in transformed space.

Preliminarily, all the samples (including test ones) are moved so that the mean of the training set becomes at the origin of the coordinate (i.e., $\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)} = 0$).

Data whitening transformation

Let $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ be the eigenvalues matrix and eigenvectors matrix of sample estimate of simplified covariance matrix $\mathbf{S}_\mathrm{S}$, i.e., $\mathbf{S}_\mathrm{S} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T$. All the test and training samples are transformed by $\boldsymbol{y} = \boldsymbol{\Lambda}^{-½}\boldsymbol{\Phi}^T\boldsymbol{x}$. This transformation makes zero valued weights be good initial ones for subsequent learning of SLP.

Learning of SLP

Let the initial weight be zero vector. Then, the perceptron is trained by gradient descent method. After the first batch iteration, we obtain discriminant function $g(\boldsymbol{y}) = \left(\boldsymbol{y} - ½\left(\bar{\boldsymbol{y}}^{(1)} + \bar{\boldsymbol{y}}^{(2)} \right) \right)^T \left(\bar{\boldsymbol{y}}^{(1)} - \bar{\boldsymbol{y}}^{(2)} \right) k_\mathrm{E}$, where $\bar{\boldsymbol{y}}^{(1)}$ and $\bar{\boldsymbol{y}}^{(2)}$ are the sample mean vectors in the transformed space, and k_E is a scalar constant. This DF is equal to transformed DF (1), i.e. $g(\boldsymbol{y}) = \left(\boldsymbol{y} - ½\left(\bar{\boldsymbol{y}}^{(1)} + \bar{\boldsymbol{y}}^{(2)} \right) \right)^T \left(\bar{\boldsymbol{y}}^{(1)} - \bar{\boldsymbol{y}}^{(2)} \right)$.

The data whitening transformation gives good initial weights for training of SLP as long as the both classes share common CM and the distributions are well-estimated. For more details, see book [10]. Theoretically and practically, in subsequent training, SLP outperforms Fisher classifier if samples are non-Gaussian.

3 Structurized and Regularized Estimates of Covariance Matrix

In our research work, nine kinds of covariance matrix models are used (Fig. 4). First of all, three models without clear structurization of CM were considered. They are statistical estimators rather than structurization methods. Model **FULL** use regularized pooled CM $\mathbf{S}_{\mathrm{FULL}} = \mathbf{S}$. In model **FULL**, we consider all $p(p\text{-}1)/2$ correlations (off-diagonal elements of CM). Model **NO** uses identity matrix. Note all CM models in this section become **NO** when $\lambda_0 = 1$. Model **SQDF** is a method which bases statistical classifier SQDF [12]. Unlike to other models, class-dependent CM's are separately calculated and then pooled. Small eigenvalues of each CM are replaced by a constant. The constant is estimated by the maximum likelihood estimation. Here, the number of eigenvalues which are not constant is 5.

Based on the assumption that *most correlations between distant pixels are low*, three types of fixed structure models specialized for feature vectors of 2D image were used (see the description for the feature vector at the very beginning of Section 4). The fixed structure models used block-diagonal CM: **49B4** has 49 independent 4×4 blocks, **4B49** has 4 blocks of size 49×49, and **4B&49B** covers both regions of **49B4** and **4B49**. Because both **49B4** and **4B49** are rather restrictive ones, less restrictive and more sophisticated model, **4B&49B**, is designed.

In addition to fixed structure models, we investigated three adaptive structured models. As a "dumb" model, we employ **LARG**, where much smaller correlations of CM are ignored so that the CM becomes close to diagonal. The second one is standard first-order tree dependence model **TREE1** [8, 9]. Here, it is postulated that each feature depends only on one other feature. Therefore, an inverse of the matrix , $(\mathbf{S}_{\mathrm{TREE1}})^{-1}$, however, has $2p$-1 non-zero elements. In general case, however, the inverse of non-structurized CM has $p\times p$ non-zero elements. In the previous research studies [9], most often this model appeared as a best one in moderate sample size situations. The last one is **EBD** which block-diagonalizes CM after exchanging elements of the matrix in order that sub-covariance matrices contain larger elements [13]. 14 is used for the number of sub-matrices.

In the earlier stage of the investigation, we also considered scaled rotation regularization [10, 14]. In experiments with 196-dimensional data and relatively small learning sets (100 samples), this regularization method was too complex and ineffective.

If the number of training samples is too small, stucturized estimate of CM, $\mathbf{S}_{\mathrm{S}}$, is unreliable. For more reliable estimation, we are obliged to introduce additional regularization, i.e. $\mathbf{S}_{\mathrm{S\&RDA}} = (1-\lambda)\mathbf{S}_{\mathrm{S}} + \lambda\mathbf{I}$, where λ is a parameter for regularization. If $\lambda = 0$, we have no regularization. If $\lambda = 1$, we have no structurization (case **NO**). Intermediate values of parameter λ could improve accuracy of determination of the weight vector obtained after the first batch iteration. Because good initial weight vector leads good result if training would be stopped in a right moment, proper additional regularization should assist in reducing generalization error.

Note, if regularization is applied to conventional sample estimate of CM (in case of **FULL**), in dependence on number of iterations we have RDA

($\mathbf{S}_{\mathrm{RDA}} = (1-\lambda)\mathbf{S} + \lambda\mathbf{I}$,) with different λ (e.g., see Eq.(4.9) in [10]). RDA is known as one of the best classification methods in statistical pattern recognition.

4 Experiments

In the experiments, 196-dimensional directional element feature [15] was used to represent handwritten Japanese characters in database ETL9B. Preliminary to extracting the feature vector, a character image was normalized nonlinearly [16] to fit in a 64×64 box. Then, skeleton were extracted, and line segments of vertical, horizontal and slanted at ±45 degrees were extracted. An image is divided into 49 sub-areas of 16×16 dots (see Fig.3). Sum of each segment in a region is an element of feature vector.

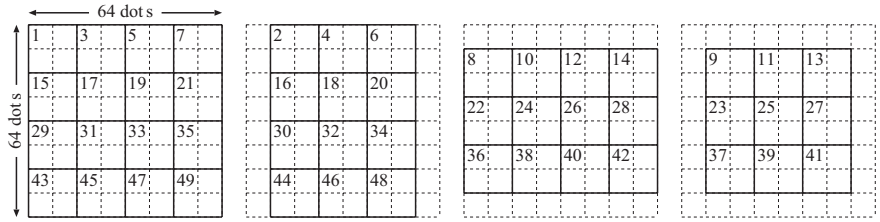


Fig. 2. 49 sub-areas of feature vector.

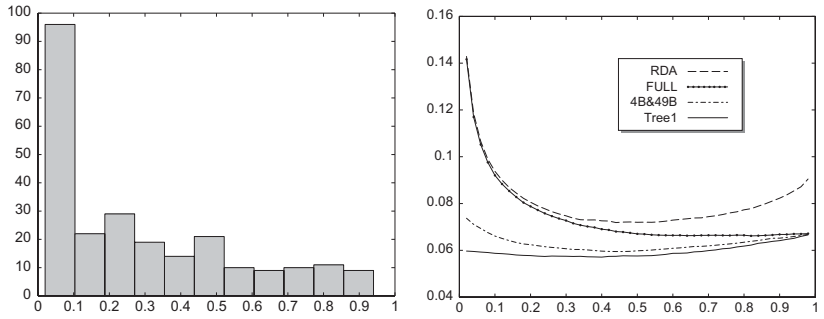
Our purpose is to investigate potential possibilities of each of pattern classification method (strategy) in very difficult case where training set size $n=N_1+N_2 \approx p$. Therefore, for 196-dimensional feature vector, we considered $N = N_1 = N_2 = 30, 50, 100, 150$. This is a really critical situation of small sample/high dimensionality problem. In each experiment, we used test sets to find optimal regularization parameter which achieves minimum error rate. Each time we permuted 200 vectors in each pattern class. In each category, N samples were selected for training and remaining 200-N ones were for testing.

Preliminary experiments demonstrated that test error estimates depend on value λ notably. Optimal values of λ depend on CM structurization method, training set size and also on random split of data into training and test sets. In Fig. 3a, we present typical histogram of distribution in 250 experiments for model **Tree1**. In Fig. 3b, we have generalization errors as function of λ for five CM structurization methods (**RDA, FULL**, **4B&49B** and **Tree1**) calculated from 250 experiments.

We analysed bivariate distributions of optimal values in Fig.3a. We found there is no or very small correlations between two distinct CM models considered. This means for each CM structurization model, one needs utilize its own (best) value of λ. Accordingly, optimal λ is CM structurization method dependent. For this reason, for each pair and CM structurization method for all handwritten character pairs in Fig.1 (named as **A** to **N** from upper left pair), we performed ten preliminary experiments to evaluate approximately a fixed value of optimal λ to be used in the main experiments.

Average results obtained in 100 experiments for character pairs are presented in Table 1. We see that joint utilization of prior information in form of postulated structure and additional regularization of CM are useful even when parameter λ is determined approximately. We found that there is no single CM structure best for all handprinted character pairs. Most often, fixed structure models such as **49B4**, **4B49** and **4B&49B** were the best. In several cases, statistical structure models such as **Tree1** and **SQDF** outperformed the fixed structure models.

Experiments with different training set sizes are shown in Table 2. This also confirmed usefulness of joint utilization of the CM structurization and regularization. The generalization error decreases uniformly with training set size N. The best two CM structurization models do not change with an increase in training set size.



(a) Distribution of optimal λ (b) Generalization errors as function of λ

Fig. 3. Optimal λ and generalization errors (Pair **M**, N=100, in 250 experiments): (a) distribution of optimal regularization parameter λ for model **Tree1**, and (b) generalization errors as functions of λ for **RDA**, **FULL**, **4B&49B** and **Tree1**.

Table 1. Average generalization errors for different character pairs of **FULL**, and relative ratios of generalization error of each method to generalization error of **FULL** (right 9 columns). The very last rows in the table are average values of the column.

Pair	FULL	NO	SQDF	49B4	4B49	4B&49B	LARG	TREE1	EBD	RDA
A	0.1955	0.9974	**0.9514**	0.9974	1	1.0051	0.9949	0.9923	0.9974	1.0691
B	0.1475	1.0034	0.9695	0.9864	0.9966	0.9831	1.0102	**0.9322**	1.0034	1.0508
C	0.0900	1.0056	1	0.9944	1	1	0.9944	**0.9889**	1.0056	1.1556
D	0.1405	0.9929	1.0036	**0.9893**	**0.9893**	**0.9893**	1.0071	1.0036	0.9929	1.0747
E	0.1465	1.0034	1.0137	1	1.0102	1.0068	1.0171	1.0068	**0.9863**	1.0819
F	0.0410	1.0366	**0.9390**	1.0244	0.9634	0.9756	0.9634	0.9756	1.0122	1.0854
G	0.0690	1.0290	**0.9783**	1.0290	1	1.0072	1.0145	1	1.0217	1.0217
H	0.0810	1.0556	1.0309	1.0556	**0.9506**	0.9815	1.0556	0.9753	1.0617	1.0864
I	0.1515	1.0033	1.0033	0.9967	**0.8515**	0.8746	0.967	0.9703	0.9934	1.0627
J	0.0725	1.0138	0.9862	0.9034	0.9862	**0.9241**	0.9724	1.0069	1.0138	1.0828
K	0.0950	0.9895	0.9842	0.9368	0.9579	0.9211	0.9211	**0.9158**	0.9632	1.0368
L	0.0785	1.0191	**0.9554**	1.0191	1.0127	1.0127	1.0255	1.0127	0.9873	1.1274
M	0.0640	1.0391	1.0156	1.0078	0.9531	0.9531	0.8828	**0.7891**	1.0313	1.1172
N	0.0450	1.0111	0.9889	**0.9778**	1	1	1.0111	1.0111	0.9889	1.0667
Mean	0.1012	1.0143	0.9871	0.9942	0.9765	0.9739	0.9884	**0.9700**	1.0042	1.0799

Table 2. Average generalization errors for different training set sizes, *N*, and diverse CM structurization methods (Pair **M**).

N	FULL	NO	SQDF	49B4	4B49	4B&49B	LARG	TREE1	EBD	RDA
30	0.1151	0.1145	0.1146	0.1118	0.1089	0.1052	0.1116	**0.0965**	0.1145	0.1250
50	0.0895	0.0902	0.0914	0.0883	0.0846	0.0820	0.0865	**0.0769**	0.0904	0.0980
100	0.0624	0.0646	0.0658	0.0633	0.0595	0.0582	0.0614	**0.0565**	0.0648	0.0691
150	0.0537	0.0543	0.0549	0.0515	0.0491	0.0487	0.0490	**0.0452**	0.0550	0.0608

5 Concluding remarks

In the current paper, we considered performance of the integrated approach of statistical estimators and neural networks. The main purpose is to investigate potential possibilities of this approach combined with various strategies under very difficult condition where the number of sum of training vectors is almost dimensionality. We used similar pairs of handwritten Japanese characters. This aggravates more difficult situation.

The strategies we used were 1) utilization of prior information in form of postulated structure of covariance matrix; 2) regularization of CM; 3) solution of the perceptron is regularized by early stopping before a minimum of the cost function. As prior information, nine kinds of structurization methods (models) were used. They also could be grouped as statistical models, fixed structure ones and adaptive structure ones. Fixed structure models are designed for feature vector of 2D spatial image. Regularization of CM directly improves data transformation which gives initialization of the perceptron. The number of learning steps of SLP decides complexity of pattern classification algorithm, too.

In experiments, utilization of structure models allowed us to reduce generalization error for most of the character pairs. For all 14 character pairs considered, error of "the best method" was on average 1.15 times smaller in comparison with the optimized regularized discriminant analysis. It was 1.05 times smaller than that of SLP with regularized maximum likelihood covariance matrix (model **FULL**) utilized for preliminary data transformation. Results of our research pointed out that joint utilization of structurization and conventional regularization of CM has a potential to improve efficacy of the integrated approach in designing pattern classifiers. The experiments show no structure is the best for all pairs. Therefore, the best structure and regularization parameter have to be selected for each pair and sample size respectively.

All three regularization techniques are acting simultaneously in the same directions. Thus, each of them can influence (reduce) effectiveness of other two. The effects of CM structures were also aggravated by the fact that most of distributions of the input features are highly asymmetric or bimodal, i.e. assumptions about Gaussian distributions were violated markedly [17]. In future research, the effects of factors aggravating positive effects have to be considered in detail. Practical techniques to select proper values of regularization parameters and optimal iteration should be developed.

Acknowledgments

The authors thank Assoc. Prof. Shinichiro Omachi, Prof. Hirotomo Aso, Dr. Ausra Saudargiene and Giedrius Misiukas for useful discussions, shared with us data sets and Matlab codes.

References

1. D. Morgera, D.B. Cooper. Structurized estimation: Sample size reduction for adaptive pattern classification. *IEEE Trans. Information Theory*, 23:728-741,1977.
2. V. Kligys. On the classification of multivariate Markov sequences. *Statistical Problems of Control*, Inst. of Math. and Cyb. Press, Vilnius, (S. Raudys, ed.), 50:57-75, 1981 (in Russian).
3. D.A. Landgrebe. The development of a spectral-spatial classifier for earth observational data. *Pattern Recognition*, Vol. 12:185-175, 1980.
4. D. Morgera. Linear, structured covariance estimation: An application to pattern classification for remote sensing. *Pattern Recognition Letters*, 4(1): 1-7, 1986.
5. G. Palubinskas. Spatial image recognition. *Statistical Problems of Control*, Inst. of Math. and Cyb. Press, Vilnius, (S. Raudys, ed.), 74:104-113, 1986 (in Russian).
6. G. Palubinskas. A comparative study of decision making algorithms in images modeled by Gaussian random fields. *Int. J. of Pattern Recognition and Artificial Intelligenc*e. Vol. 2(4):621-639, 1988.
7. G. Palubinskas. A review of spatial image recognition methods. *Statistical Problems of Control*, Inst. of Math. and Cyb. Press, Vilnius, (Raudys S., ed.), 93:215-231, 1990 (in Russian).
8. S. Raudys, A. Saudargiene. Structures of the covariance matrices in the classifier design. *Lecture Notes in Computer Science*, Springer-Verlag, 1451:583–592, 1998.
9. S. Raudys, A. Saudargiene. Tree type dependency model and sample size - dimensionality properties. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 23(2):233-239, 2001.
10. S. Raudys. *Statistical and Neural Classifiers: An integrated approach to design*. Springer, NY, 2001.
11. S. Raudys, S. Amari. Effect of initial values in simple perception. *Proceedings 1998 IEEE World Congress on Computational Intelligence, IJCNN'98*, 1530-1535, 1998.
12. S. Omachi, F. Sun, H. Aso. A new approximation method of the quadratic discriminant function. *Lecture Notes in Computer Science*, 1876: 601-610, 2000.
13. F. Sun, S. Omachi, N. Kato, H. Aso, S. Kono, T. Takagi. Two-stage computational cost reduction algorithm based on Mahalanobis distance approximations. *Proceedings 15th Int. Conf. on Pattern Recognition* (ICPR2000), IEEE Press, 2:700-703, 2000.
14. S. Raudys. Scaled rotation regularization. *Pattern Recognition* 33:1989–1998, 2000.
15. N. Sun, Y. Uchiyama, H. Ichimura, H. Aso, M. Kimura: Intelligent recognition of characters using associative matching technique. *Proc. Pacific Rim Int'l Conf. Artificial Intelligence* (PRICAI'90), 546-551, 1990.
16. H. Yamada, K. Yamamoto, T. Saito. A nonlinear normalization method for handprinted kanji character recognition - line density equalization. *Pattern Recognition*, 23(9):1023-1029, 1990.
17. S. Raudys, M. Iwamura. Multiple classifiers system for reducing influences of atypical observations. *Lecture Notes in Computer Science* (Proceedings of *Multiple Classification Systems*; MCS 2004).

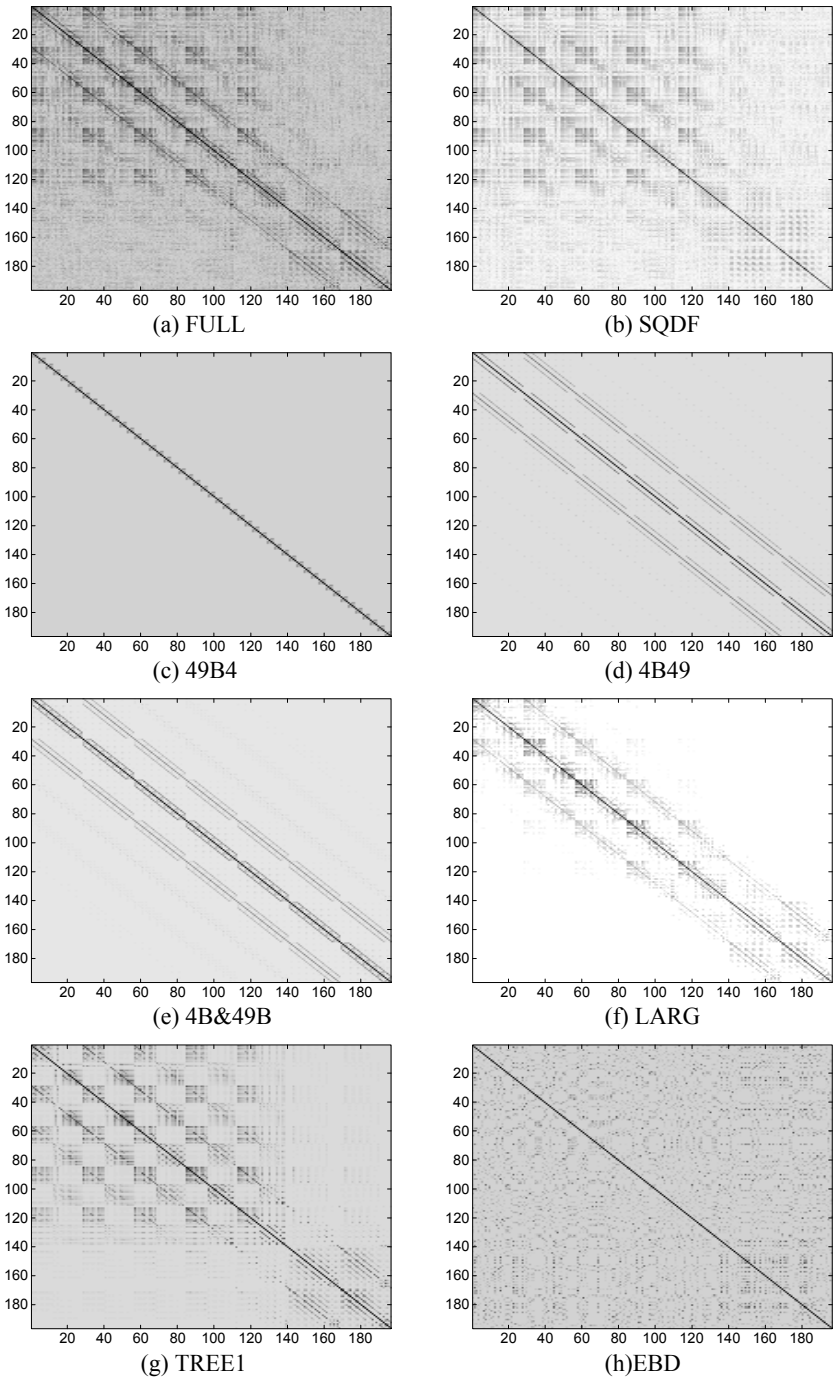


Fig. 4. Elements of structurized covariance matrices except model **NO**. Darker pixel stands for larger absolute value.