MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Regularized Covariance Matrix Estimation with High Dimensional Data for Supervised Anomaly Detection Problems

Nikovski, Daniel N.; Byadarhaly, Kiran

TR2016-099 July 27, 2016

Abstract

We address the problem of estimating highdimensional covariance matrices (CM) for the explicit purpose of supervised anomaly detection, in the case when the number n of data points is lower than their dimensionality p. This is increasingly common with the emergence of the Internet of Things that makes it possible to collect data from many sensors simultaneously, resulting in very high-dimensional data points. When we attempt to perform anomaly detection for such data by modeling the normal behavior of the system by means of a multivariate Gaussian distribution, and n < p, the sample CM is singular, and cannot be used directly without some form of regularization. In contrast to existing methods for CM regularization that aim to fit the training data accurately, we propose a regularization algorithm for CM estimation that directly aims to maximize the area under the resulting receiveroperator characteristic (AUROC) for the ultimate decision problem that needs to be solved: anomaly detection. Experiments on test problems demonstrate the ability of the proposed algorithm to find CM estimates significantly better at anomaly detection than existing estimation methods that are unaware of the decision task that the CMs they produce will be used in.

IEEE International Joint Conference on Neural Networks (IJCNN)

Regularized Covariance Matrix Estimation with High Dimensional Data for Supervised Anomaly Detection Problems

Daniel Nikovski, *Member, IEEE,* and Kiran Byadarhaly

***Abstract*—We address the problem of estimating high-dimensional covariance matrices (CM) for the explicit purpose of supervised anomaly detection, in the case when the number n of data points is lower than their dimensionality p. This is increasingly common with the emergence of the Internet of Things that makes it possible to collect data from many sensors simultaneously, resulting in very high-dimensional data points. When we attempt to perform anomaly detection for such data by modeling the normal behavior of the system by means of a multivariate Gaussian distribution, and $n < p$, the sample CM is singular, and cannot be used directly without some form of regularization. In contrast to existing methods for CM regularization that aim to fit the training data accurately, we propose a regularization algorithm for CM estimation that directly aims to maximize the area under the resulting receiver-operator characteristic (AUROC) for the ultimate decision problem that needs to be solved: anomaly detection. Experiments on test problems demonstrate the ability of the proposed algorithm to find CM estimates significantly better at anomaly detection than existing estimation methods that are unaware of the decision task that the CMs they produce will be used in.**

I. Introduction

Anomaly detection is one of the main classes of problems addressed by data mining methods. One popular general approach of such methods is to characterize the normal behavior of the system or process under observation by means of a model learned from a collected data set of examples of normal behavior. Once such a model has been created, new data points are continuously tested against the model to decide whether they are likely to conform to it, or not; if not, an alert for a possible anomaly is raised.

For many problems, a parametric model of normal behavior is appropriate, and one of the most popular parametric models is the multivariate Gaussian model $N(\mu,\Sigma)$ with mean μ and CM Σ. For example, when the operating point of a process or device is intended to be fixed, but there are multiple (possibly correlated) disturbances and/or measurement errors, by virtue of the central limit theorem, the vector of observations of that operating point is likely to be a random multivariate Gaussian variable. The probability density function (pdf) of a multivariate Gaussian distribution in p dimensions is given by

$$f(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \qquad \text{(I.1)}$$

Daniel Nikovski and Kiran Byadarhaly are with Mitsubishi Electric Research Laboratories, Cambridge, MA 02139, USA (email: nikovski@merl.com).

Here x is a p-variate Gaussian random variable, $\mu \in \mathbb{R}^p$ is the mean of the Gaussian distribution, and $\Sigma \in \mathbb{R}^{p\times p}$ is the non-negative definite CM of the Gaussian distribution. Frequently, for a given multivariate Gaussian model $N(\mu,\Sigma)$ and a test point x, the density $f(x;\mu,\Sigma)$ at that point is used as an anomaly score: if that density is below a given threshold, an anomaly is signaled.

Clearly, if the density $f(x;\mu,\Sigma)$ is to be evaluated, the CM Σ must be invertible. However, in many data-driven anomaly detection problems, this is not necessarily always the case. In such problems, a data set $X = \{x_1, x_2, x_3, \ldots, x_n\}$ with n samples has usually been collected, and Σ must be estimated from X. The maximum likelihood (ML) estimator of Σ, also known as the sample covariance matrix Σ_S, is $\Sigma_S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_S)(x_i-\mu_S)^T$, where $\mu_S = \frac{1}{n}\sum_{i=1}^{n}x_i$ is the sample mean. In high-dimensional problems, when the number of dimensions p exceeds the number of samples n, the sample CM Σ_S will be singular, and hence non-invertible; its direct use for anomaly detection by means of the multivariate Gaussian density would thus be impossible.

In this paper, we address the problem of estimating the CM Σ from a data set X in the regime $n < p$, specifically for the purpose of anomaly detection. The regime $n < p$ is becoming increasingly important in practical applications, with the advent of the Internet of Things, itself facilitated by increasingly affordable sensing, communication, and information technologies. Modern machines and entire systems are increasingly equipped with multiple sensors that produce one or more senor readings, aggregated into a joint high-dimensional vector x that describes the state of the entire system. In many cases, it is desirable to produce reasonable and usable model estimates from relatively few collected data samples, in comparison to the dimensionality of the data vector; this is when estimation in the regime $n < p$ is often necessary.

Due to the high practical significance of this regime, a lot of research has been focused on it in the statistical and machine learning communities. A number of advanced methods, such as Hoffbeck and Ledoit-Wolf regularization, as well as banding and tapering algorithms, have been proposed, and are briefly reviewed in Section II. The main commonality between these methods is that they approach the problem as a data fitting, machine learning problem, whereas the problem of anomaly detection is essentially a *decision* problem. The objective of these methods is to optimize a loss function that describes how well the estimated CM fits the data set X, while still providing a well-behaved, non-singular matrix. For example, Hoffbeck

regularization maximizes the log-likelihood of the data set X, and Ledoit-Wolf regularization minimizes the Frobenius norm of the difference between the estimated and true CMs. Although these approaches have good statistical merits as regards the accuracy of fitting the data set X, that accuracy is not the ultimate goal of anomaly detection algorithms — instead, the ultimate goal is high accuracy of the decision problem, namely anomaly detection. That accuracy is typically measured by different loss functions, for example the area under receiver operator characteristic (ROC) curves. In Section III, we propose a novel algorithm that directly optimizes such a loss function for the purposes of anomaly detection, and in Section IV, we investigate its performance on test problems, and compare it to that of methods that are concerned with optimizing purely data fitting measures.

II. CM Estimation Methods in the Regime $n < p$

For high-dimensional data, unless the number of samples is much larger than the number of dimensions, for example, $n > 10p$, an accurate estimation of the covariance matrix is not likely. In such cases, one of the simplest approaches is to use the diagonal CM Σ_d in place of the full CM. Although it can typically be estimated from very few samples, its use would completely ignore the cross correlations between pairs of individual variables, and is not likely to result in high accuracy when $2 \ll n < p$ [1].

A much more successful approach to CM estimation when $n < p$ is to regularize the singular sample CM by blending it with another positive definite matrix of full rank. These methods are also known as shrinkage methods. Given the singular sample CM Σ_S estimated from insufficient data, the regularized covariance Σ_R is given by

$$\Sigma_R = (1-\alpha)\Sigma_S + \alpha\Sigma_F \tag{II.1}$$

Here α is the shrinkage parameter, and Σ_F is a chosen full rank matrix; the original low-rank sample CM is essentially "shrunk" to that matrix, hence the name of this class of methods. The full-rank CM can either be Σ_d, or the identity matrix I_p scaled by the trace σ of the full sample CM, or any other suitable full-rank matrix. Depending on the type of that matrix, and the method used to determine the shrinkage parameter α, several methods have been proposed, as described below.

A. Hoffbeck Regularization

An effective regularization method for covariance matrix estimation proposed by Hoffbeck et al. [2],[3] uses leave-one-out cross-validation (LOOCV) on a training set and maximizes the likelihood of data left out in a computationally efficient way. In this method, the estimated low rank covariance matrix Σ_S is shrunk towards the diagonal covariance matrix Σ_d. In LOOCV, each training sample is omitted in turn, then the sample CM is estimated from the remaining samples, after which the log-likelihood of the omitted sample is computed with respect to the estimated mean and the regularized covariance of the remaining data, for a given choice of regularization parameter α. The log-likelihood of omitted samples is averaged over the entire data set, and the value of α with the highest average log-likelihood is selected. The final regularized covariance is the combination of the full and diagonal covariances estimated over the entire training set.

B. Ledoit-Wolf Regularization

Ledoit and Wolf [4] proposed a method for the estimation of covariance matrices in which they compute a well-conditioned structured estimator and then shrink the sample covariance to that structured estimator. The well-conditioned structured estimator used is the identity matrix scaled by the trace of the original sample covariance matrix. The optimal combination of the sample covariance and the structured estimator is obtained by minimizing the Frobenius norm of the difference between the regularized and true CM.

The main difficulty in obtaining accurate covariance estimation in the Ledoit-Wolf method is that the optimal combination of the sample covariance and the structured estimator depends on the true covariance matrix which is not known. To resolve this problem, the authors proposed a consistent estimator of the parameters involved in the computation of the optimal shrinkage parameter, and proved that substituting these consistent estimators for the true parameters does not make any difference asymptotically. The problem of minimizing the Frobenius norm of the difference between the true covariance and the estimated regularized covariance can be written as the following quadratic programming problem under an equality constraint:

$$\begin{aligned} \underset{\alpha}{\text{minimize}} \quad & E[\| \Sigma_R - \Sigma \|^2] \\ \text{subject to} \quad & \Sigma_R = (1-\alpha)\Sigma_S + \alpha\sigma I \end{aligned} \tag{II.2}$$

Here Σ is the true covariance, and σ is the trace of the sample covariance matrix Σ_S. The optimal solution to the above problem can be obtained analytically as $\alpha^* = \beta^2/\delta^2$, where $\delta^2 = E[\|\Sigma_S - \rho I\|^2]$, $\rho = \langle \Sigma, I \rangle$, and $\beta^2 = E[\|\Sigma_S - \Sigma\|^2]$. Clearly, these quantities cannot be computed directly without prior knowledge of the true CM Σ, but they can be replaced with their readily computable consistent estimators $d^2 = \| \Sigma_S - rI \|^2$, $r = \langle \Sigma_S, I \rangle$, and $b^2 = min(\bar{b}^2, d^2)$, where $\bar{b}^2 = \frac{1}{n^2}\sum_{k=1}^{n} \| x_k (x_k)^t - \Sigma_S \|$. When these consistent estimators are replaced in the expression for the regularization parameter α, its optimal value for the Ledoit-Wolf method is obtained as $\alpha_{LW} = b^2/d^2$. In general, this value will be different from that obtained by Hoffbeck's method.

C. Band-Diagonal Covariance

Another method proposed by Bickel and Levina uses a banded version of the sample covariance matrix [1]. They show that by banding the covariance matrix, they can get consistent and non-singular estimates of the CM under the l_1 norm. For a given full sample covariance matrix Σ_S with individual entries σ_{ij}, and a band limit k, $0 \leq k < p$, its banded matrix $\Sigma_k = B_k(\Sigma_S)$ is given by $\Sigma_k \doteq [\sigma_{ij} 1(i-j \leq k)]$, where $1(\cdot)$ denotes the indicator function that is equal to 1

when its argument is true. The most important parameter here is the choice of the banding parameter k. Ideally, it would be chosen such that the norm of the difference between the banded covariance matrix and the true covariance matrix Σ is as small as possible; however, again, the true covariance matrix is not available. Instead, the authors proposed to divide the data set into two parts, producing two sample covariance matrices Σ_{S_1} and Σ_{S_2}. If we define a risk variable $R_B(k) \doteq \|B_k(\Sigma_{S_1}) - \Sigma_{S_2}\|_{(1,1)}$, we choose the optimal band width k^* such that $k^* = \underset{k}{argmin} R_B(k)$. The matrix norm used is the l_1 matrix norm.

D. Band-Tapering Covariance

A tapering procedure for the estimation of the covariance matrix was proposed by Cai, Zhang, and Zhou [5]. For a given integer k with $1 \leq k \leq p$, a tapering estimator of the covariance matrix is defined as the weighted modification $\Sigma_T = T_{k,\lambda}(\Sigma_S) = (w_{ij}\sigma_{ij})_{p\times p}$ of the entries σ_{ij} of the sample covariance matrix Σ_S, with weights given by

$$w_{ij} = \begin{cases} 1, & when\, |i-j| \leq k_h \\ 2 - \frac{|i-j|}{k_h}, & when\, k_h < |i-j| < k \\ 0, & otherwise \end{cases} \qquad \text{(II.3)}$$

Here, $k_h = k/2$. The authors have chosen an optimal trade-off value for $k = n^{\frac{1}{2\lambda+1}}$ for which they find upper and lower bounds. The value of the parameter λ must be chosen so as to minimize the norm between the tapering covariance matrix and the true covariance matrix. Similarly to the banded case, a risk $R_T(\lambda) \doteq \|T_{k,\lambda}(\Sigma_{S_1}) - \Sigma_{S_2}\|_{(1,1)}$ is defined for a split of the data set, and the optimal parameter λ^* is chosen as $\lambda^* = \underset{\lambda}{argmin} R_T(\lambda)$.

III. Regularized CM Estimation for Anomaly Detection

All methods described in the previous section aim to optimize the fit of the regularized CM to the data. In contrast, we are considering the problem of estimating useful regularized CMs for the decision problem of supervised anomaly detection, which calls for a different optimization criterion that is specific to the nature of the decision problem. One popular definition of anomaly detection is that it is the process of identifying data points that do not conform to a notion of expected normal behavior. The most important aspect of anomaly detection is thus to accurately quantify the notion of normal behavior. The training data that is used to learn the anomaly detection model must strongly reinforce the concept of normal behavior so that the model is trained well enough to identify anomalies in data that it has not seen before. One of the ways it can be done is if the training data has explicit labels that indicate if each point is an anomaly or not. These labels must be provided by the experts who have seen the data and know what exactly constitutes an anomalous behavior [6], [7], [8]. In many cases, such data is available from maintenance records of industrial machinery or systems.

The methods described in the previous section are often successful in producing non-singular CM estimates that fit a particular data set X well. In principle, when such CMs are used within a decision problem, such as classification, regression, etc., the estimated CMs can be used directly in the predictive model. For example, for a two-class classification problem, where the two classes have multivariate Gaussian distributions with different means, but share the same CM, the optimal classifier can be shown to be a Linear Discriminant Analysis (LDA) classifier; similarly, when the CMs for the two classes are different, the optimal classifier is a Quadratic Discriminant Analysis (QDA) classifier [9]. For a large number of classification problems, this kind of symmetric assumption — that both classes are characterized by multivariate Gaussian distributions — is largely reasonable. The decision surfaces for these classifiers would be hyperplanes (for LDA) or quadratic surfaces (for QDA).

However, for the problem of supervised anomaly detection, the statistical characteristics of normal and abnormal data points are typically very different. As argued in the beginning, when normal behavior corresponds to a fixed operating point of a system or a device that is disturbed by random measurement or process noise, it is often reasonable to assume that data corresponding to normal behavior comes from a multivariate Gaussian distribution. Conversely, this assumption is typically not reasonable for the abnormal data. Because abnormal is defined as anything that is not normal, it is not likely that abnormal data points would cluster around a specific value; rather, they are more likely to be scattered around the entire operating domain. So, when examples of abnormal operation are available, estimating a CM from such data would not make sense, because their distribution is not Gaussian; it could instead be uniform, multi-modal, etc. In this case, the optimal decision surface is likely to be a hypersphere corresponding to an isocontour of the probability (for example, at the 1% probability level) given by the Gaussian distribution of the normal class. Clearly, this decision surface is very different from those of the LDA and QDA classifiers.

As mentioned, supervised anomaly detection can be viewed as a classification problem, and any number of supervised learning algorithms can be used to solve it, for example decision trees, support vector machines (SVM), multi-layer neural networks, etc. In the case of a multi-layer neural networks, one or more hidden layers with a large number of neurons may be needed to learn decision surfaces for high dimensional nonlinearly separated data of the type described above. Similarly, SVMs with standard kernels would likely find it quite difficult to separate this type of data in high dimensions without a large number of samples. Given sufficient data, these algorithms should be able to learn a decision surface close to the optimal. However, they would not make use of the knowledge that at least the normal class has a multivariate Gaussian distribution.

Another alternative approach is to ignore the abnormal data altogether, and treat the anomaly detection problem as unsupervised, by fitting a regularized CM only for the normal

class. Although this approach should be able to produce a good anomaly detector in the asymptotic case, when a lot of data is available, its performance is likely to suffer in the $n < p$ regime, where the CM will be regularized with the only purpose of fitting the normal data, unaware of how well the resulting anomaly detector detects abnormal data points.

Below, we propose a method that makes full use of the parametric Gaussian form of the normal class, while also using the available abnormal data points to increase accuracy of anomaly detection. Its main idea is to find the optimal regularization parameter α while performing cross-validation on a hold-out set specifically for the target decision problem, anomaly detection. The objective function to be optimized in this process is not one of the usual data-fitting loss functions for other shrinkage algorithms, but the Area Under the Receiver Operator Characteristic (AUROC) that is commonly used in detection problems. Because AUROC is not differentiable with respect to the regularization α, we use a related approximating function.

In order to compute the ROC curve of an anomaly detector, a suitable anomaly score s_i is computed for every data point x_i by means of the learned predictive model. Some suitable scores are the estimated probability density $f(x_i;\mu_S,\Sigma_R)$ of the normal class evaluated at the data point x_i, or the generalized Mahalanobis distance $s_i = (x_i-\mu_S)^T\Sigma_R^{-1}(x_i-\mu_S)$ from the point x_i to the sample mean μ_S of the normal class. The AUROC can then be calculated using the Wilcox-Mann-Whitney (WMW) U statistic, which for a given labeled data set is given by [10]

$$U = \sum_{i=0}^{n^+-1}\sum_{j=0}^{n^--1} I(s_i^+, s_j^-) \tag{III.1}$$

Here s_i^+ is the score of the i_{th} data point in the abnormal class $(+)$, and s_j^- is the score of the j_{th} data point in the normal class $(-)$, n^+ is the number of examples of abnormal behavior, and n^- is the number of examples of normal behavior. The function $I(s_i^+, s_j^-)$ is given by

$$I(s_i^+, s_j^-) = \begin{cases} 1, & if\ s_i^+ > s_j^- \\ 0, & otherwise \end{cases} \tag{III.2}$$

The true AUROC can then be computed as $J_0 = \frac{U}{n^+n^-}$.

The analytic form of the AUROC, calculated using the WMW U statistic, is not differentiable and cannot easily be used as an objective function for an optimization problem. However, the ordering test $I(s_i^+, s_j^-)$ can be approximated by a sigmoid function [11]:

$$H(s_i^+, s_j^-) = \frac{1}{1+e^{-\theta(s_i^+ - s_j^-)}} \tag{III.3}$$

Here θ is a smoothing parameter for the sigmoid. A small value of the smoothing parameter θ softens the function $I(s_i^+, s_j^-)$ too much, whereas a large value of θ, although approximating $I(s_i^+, s_j^-)$ closely, would lead to numerical problems during optimization due to steep gradients.

The approximated AUROC can then be written as

$$J = \frac{\sum_{i=1}^{n^+}\sum_{j=1}^{n^-} H(s_i^+, s_j^-)}{n^+n^-} \tag{III.4}$$

The objective function (approximated AUROC) can be rewritten as a function of the regularization parameter α by expressing the normal and anomaly scores in terms of the generalized Mahalanobis distance form, as shown below:

$$\begin{aligned} J(\alpha) &= \frac{\sum_{i=1}^{n^+}\sum_{j=1}^{n^-} H[(x_i^+-\mu_S)^T\Sigma_R^{-1}(x_i^+-\mu_S), (x_j^--\mu_S)^T\Sigma_R^{-1}(x_j^--\mu_S)]}{n^+n^-} \\ &= \frac{\sum_{i=1}^{n^+}\sum_{j=1}^{n^-} H\{(x_i^+-\mu_S)^T[(1-\alpha)\Sigma_S+\alpha\Sigma_d]^{-1}(x_i^+-\mu_S), (x_j^--\mu_S)^T[(1-\alpha)\Sigma_S+\alpha\Sigma_d]^{-1}(x_j^--\mu_S)\}}{n^+n^-} \\ &= \frac{1}{n^+n^-}\frac{\sum_{i=1}^{n^+}\sum_{j=1}^{n^-} 1}{(1+e^{-\{\theta[(x_i^+-\mu_S)^T[(1-\alpha)\Sigma_S+\alpha\Sigma_d]^{-1}(x_i^+-\mu_S)-(x_j^--\mu_S)^T[(1-\alpha)\Sigma_S+\alpha\Sigma_d]^{-1}(x_j^--\mu_S)]\}})} \end{aligned} \tag{III.5}$$

$$\alpha_A = \underset{\alpha}{argmax} J(\alpha) \tag{III.6}$$

$$\bigtriangledown J(\alpha) = \frac{\sum_{i=1}^{n^+}\sum_{j=1}^{n^-} H(s_i^+-s_j^-)(1-H(s_i^+-s_j^-))\theta[(x_i^+-\mu_S)^T[\Sigma_R^{-1}[\Sigma_S-\Sigma_d]\Sigma_R^{-1}](x_i^+-\mu_S)-(x_j^--\mu_S)^T[\Sigma_R^{-1}[\Sigma_S-\Sigma_d]\Sigma_R^{-1}](x_j^--\mu_S)}{n^+n^-} \tag{III.7}$$

The optimal value α_A of the regularization parameter given by equation III.6 can be found by gradient ascent optimization. The gradient of the objective function with respect to the regularization factor is given by Equation III.7. We will refer to this method of finding the optimal value of α as the Area Under ROC curve REGularization (AUROCREG) algorithm.

IV. Empirical Evaluation

A. Evaluation on Synthetic Data

We compared the performance of various covariance estimation methods listed in the previous sections by using them in anomaly detection tasks on both synthetic and real test data. The multivariate Gaussian model for anomaly detection with the different estimated covariances were also compared with a kernel SVM classifier with a radial basis function kernel, as well as with a multilayer neural network (MLP) with a single hidden layer implementing the anomaly detection task as a supervised binary classification task.

Synthetic datasets containing normal and anomalous data were generated for dimensions $p \in \{20, 50, 100\}$. A total of 10,000 data samples (5,000 normal and 5,000 abnormal) were generated in each dataset. In each of these datasets, the normal data points were drawn from a multivariate Gaussian distribution with a chosen mean and covariance matrix, whereas the anomalous data points were drawn from a uniform distribution in a certain range such that it surrounds the normal data. The covariance of normal data was chosen as a random positive definite matrix. The mean of the normal data was chosen as $\mu = 0$, a p-dimensional vector of zeros.

The data generated was split into training, cross-validation, and testing sets. The number of samples in the training data set n (which is made up of only normal samples), was varied, and CMs were estimated from that data set according to each tested method. Such a training data set, made up of only normal samples for the multivariate Gaussian model, corresponds to a situation where a relatively small amount of data samples are acquired from a tested system during its normal operation. For the AUROCREG method, the cross-validation data set was used to find the optimal value of the regularization parameter α. The estimated covariances were then used for anomaly detection on the test data set, as described above.

The cross-validation and the test data set were made up of equal number of normal and abnormal samples. The training and cross-validation data points were chosen randomly from the first half of the normal and abnormal data (5,000 samples), such that when the training data consisted of n normal samples, the corresponding cross-validation set consisted of n normal and n abnormal data points. The training data for the SVM and the multilayer perceptron neural network was made up of $n/2$ normal and $n/2$ abnormal data points from the first half of the normal and abnormal data (5,000 samples). The test data was chosen randomly from the second half of the normal and abnormal data (also 5,000 samples), and always had the same size: 250 normal and 250 abnormal data points, for the sake of accurate evaluation of the resulting AUROC for anomaly detection. This process was repeated for r trials, and the resulting AUROC curves were computed from the aggregated anomaly scores over all r trials.

The first set of results examine the cost function that is being maximized to find the optimal regularization parameter α in the AUROCREG method. Figs. 1 through 3 show the dependency of the AUROC on the regularization parameter α for anomaly detection over the cross validation data sets, averaged over $r = 10$ trials. Here regularization was performed towards the identity matrix multiplied by the trace σ of the sample CM, such that the AUROCREG estimate Σ_A was obtained as $\Sigma_A = (1-\alpha)\Sigma_S + \alpha\sigma I$. It is evident that AUROC values depend strongly on α, and finding the optimal value makes a big difference in the anomaly detection task. It is also evident that the amount of training data (n) strongly affects performance, as expected. The optimal value of α could vary significantly, depending on n, p, and their relationship. In addition, it can be seen that there is relatively little difference between the true AUROC computed by the WMW statistic (shown in solid lines), and its differentiable approximation (shown in dotted lines of the same color for the same respective n). Another favorable observation is that all of these curves appear to have a single global maximum, so most nonlinear optimization algorithms should have little difficulty in finding it quickly, if supplied with the gradient of the objective function.

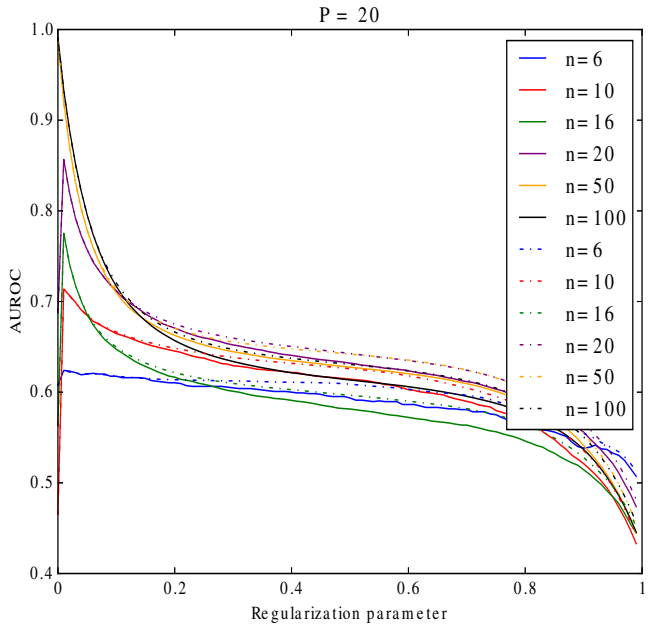

Fig. 1. Comparison between the WMW objective function and its differentiable approximation for $p = 20$ and varying values of n. The curves corresponding to the non-differentiable WMW U statistic are shown as solid lines and those corresponding to the approximate function are shown as dotted lines.

Figs. 4 through 6 show the performance on anomaly detection of multiple CM estimation algorithms (as measured by the resulting AUROC on the independent testing data set) vs. the number of available training data samples n. It can be seen that for the regime $n < p$, the AUROC achieved by the proposed algorithm AUROCREG is always at least as good, and sometimes substantially better, than that of any other regularization method tested . Only for $n = 6$ case, does some other method(s) outperform the AUROCREG algorithm,

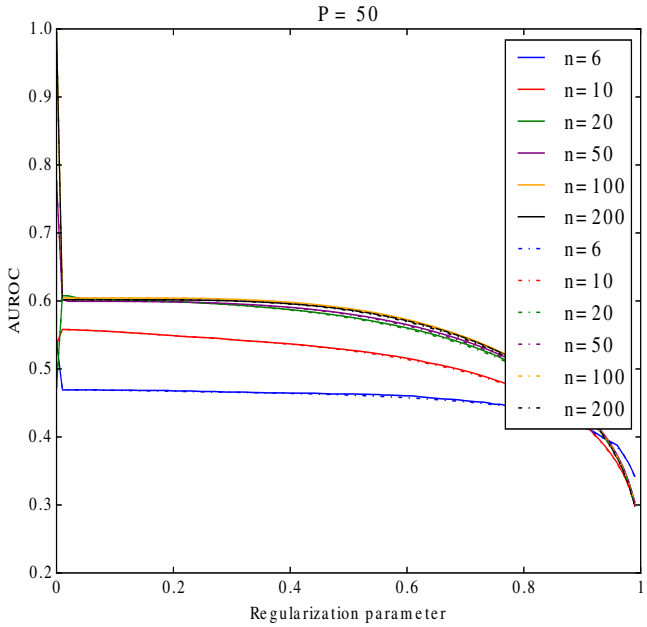

Fig. 2. Comparison between the WMW objective function and its differentiable approximation for $p = 50$ and varying values of n. The curves corresponding to the non-differentiable WMW U statistic are shown as solid lines and those corresponding to the approximate function are shown as dotted lines.

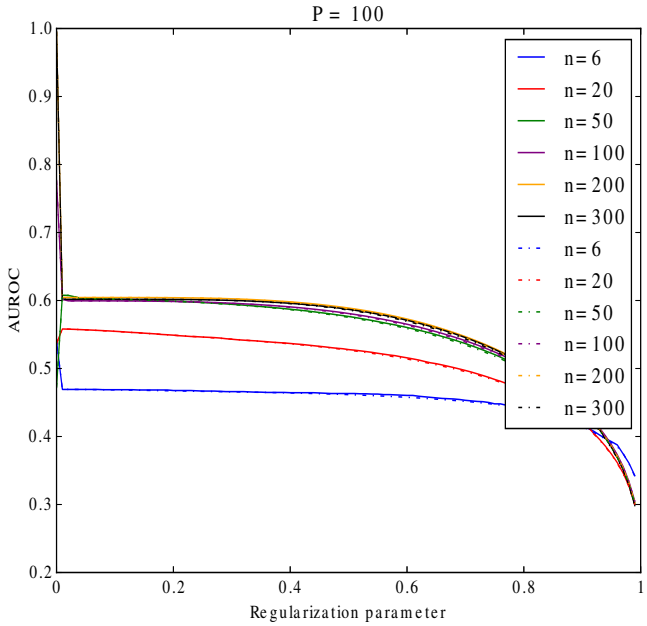

Fig. 3. Comparison between the WMW objective function and its differentiable approximation for $p = 100$ and varying values of n. The curves corresponding to the non-differentiable WMW U statistic are shown as solid lines and those corresponding to the approximate function are shown as dotted lines.

but for all others in the $n < p$ regime, the AUROCREG algorithm is consistently the best choice. For the $n \geq p$ regime, the AUROCREG algorithm clearly outperforms all other covariance estimation methods, as well as the two supervised classification methods (SVM and MLP). The band diagonal method slightly outperforms the AUROCREG method when $n = p$, but is unable to match its performance for $n > p$. This further reinforces the expectation that when the conditions are satisfied (the normal data has a Gaussian distribution), there is no better method than the multivariate Gaussian model to detect anomalies.

In these graphs, the dotted line represents the AUROC that would be achieved by the anomaly detection algorithm if it had full knowledge of the true Gaussian distribution from which the normal data samples were drawn, and thus represents the upper limit on the accuracy that any learning algorithm can achieve. It can be seen that the AUROCREG algorithm indeed approaches that accuracy in the regime $n > p$, whereas many other CM estimation algorithms fail to do that. The most important conclusion from these figures is that by optimizing directly for the performance in the task at hand (anomaly detection) rather than for data fitting measures such as log-likelihood of the training data (Hoffbeck method) or deviation from the true covariance (Ledoit-Wolf method), the performance of the anomaly detector can be increased substantially.

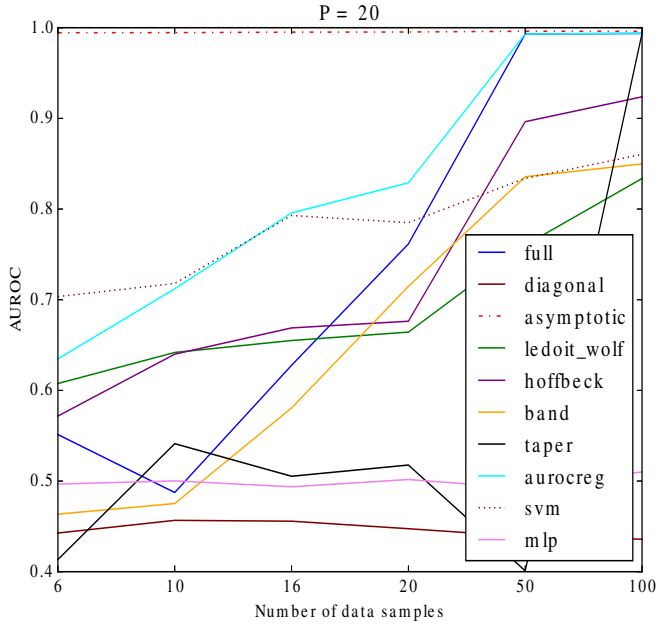

Fig. 4. Performance on anomaly detection vs. number of training data samples for $p = 20$.

B. Evaluation on Real Data

We next tested the described methods on an anomaly detection task involving vehicle silhouettes, using available

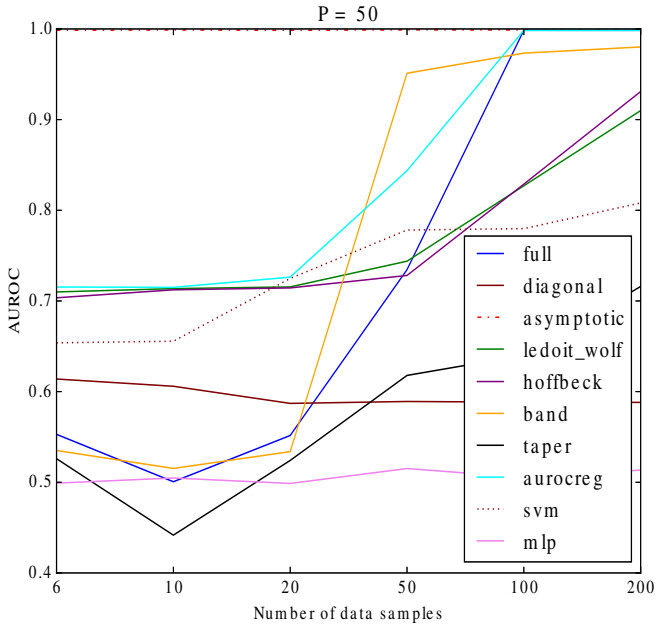

Fig. 5. Performance on anomaly detection vs. number of training data samples for $p = 50$

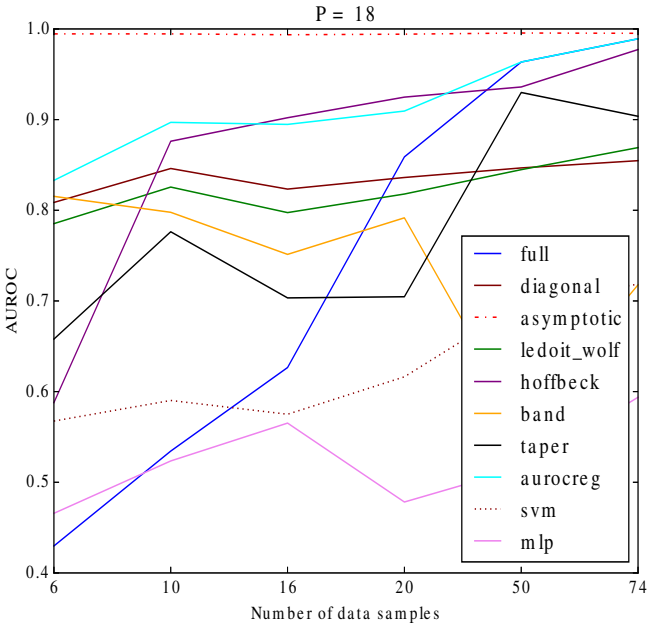

Fig. 7. Variation of the AUROC vs. the number of training data samples for the vehicle silhouette data. The number of dimensions of the data points is $p = 18$.

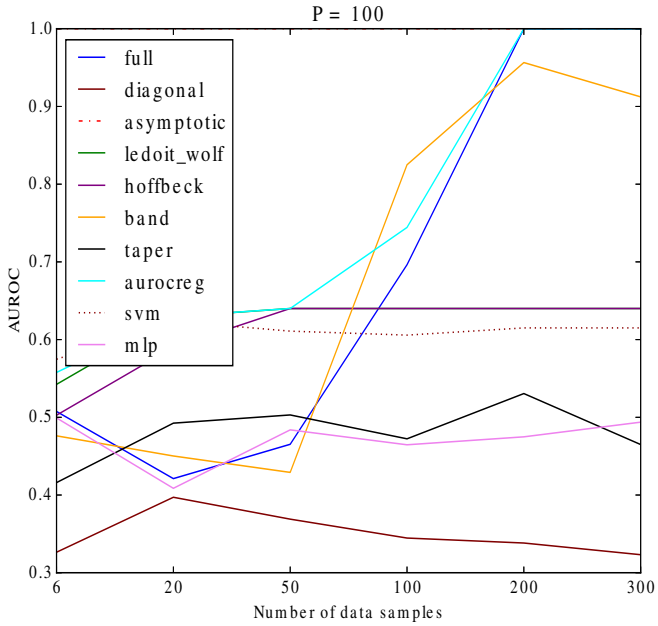

Fig. 6. Performance on anomaly detection vs. number of training data samples for $p = 100$.

data described in [12]. The purpose of this data set is to distinguish 3D objects within a 2D image based on a vector of shape features extracted from the silhouettes, of dimensionality $p = 18$. Four different model vehicles were used for the experiment, with the expectation that they would be readily distinguishable based on their feature vectors. The entire data set consists of 946 total image silhouettes for four model vehicles. In order to use this data set for anomaly detection, one of these four classes was chosen as the normal class, and the other three were merged to form the abnormal data. Note that while it might be reasonable to expect that the normal class can be modeled by a multivariate Gaussian distribution, the same assumption cannot be made for the abnormal data, because it will have at least three distinct modes. Fig. 7 shows a comparison between the accuracy of the described methods for this anomaly detection task. Clearly, the AUROCREG method dominates all other methods, in all regimes with the $n = 16$ and $n = 20$ cases being an exception in which the Hoffbeck method outperforms the AUROCREG method by a slight margin, and its performance matches that of the full sample covariance matrix Σ_S when $n \gg p$, as expected; for that regime, AUROCREG essentially uses Σ_S, by concluding that the optimal regularization parameter is $\alpha = 0$.

The described methods were also tested on a social media dataset provided by the AMG group at the Laboratoire d'Informatique de Grenoble [13]. This dataset contains data from two social networking sites, namely twitter and tom's hardware. In twitter data, there is no direct audience estimator, and they used the nad feature as the target feature and showed higher reactivity of exchanges than tom's hardware. There are French and German contributions in twitter data. In this study, they focused on a set of 6671 topics, such as: overclocking; grafikkarten; disque dur; android; etc. related to the technology domain. For the classification/anomaly detection

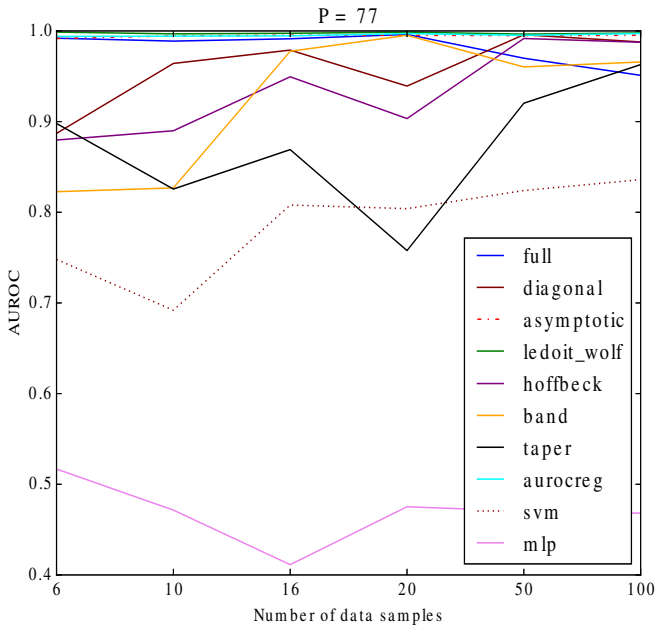

Fig. 8. Variation of the AUROC vs. the number of training data samples for the twitter data. The number of dimensions of the data points is $p = 77$.

task, they provided time windows with an upward trend in the number of tweets about a topic, and the task was to determine whether or not these time windows are followed by buzz events. The labeling and the upward change detection were done considering a univariate time-series. They also used a threshold value ≥ 500 to determine if a topic qualifies as a buzz or not. The data points that generated buzz were considered to be anomalous points, and those that didn't were considered to be normal.

Fig. 8 shows a comparison between the accuracy of the described methods for the anomaly detection task involving the social network data. The result shows that a number of methods, including AUROCREG, perform very well on the test data. Notable is the relatively poor performance of the supervised learning methods, SVM and MLP, which can be explained with the relatively high dimensionality of the feature space.

V. Conclusions and Future Work

In this paper, we have proposed a novel method for learning a well-behaved, non-singular estimate of the covariance matrix of a multivariate Gaussian distribution from limited amounts of data in high-dimensional feature spaces, specifically for the purpose of supervised anomaly detection tasks. Unlike other regularization methods that focus solely on fitting the available training data well, our method optimizes directly the expected accuracy on the anomaly detection task, as measured in the resulting AUROC. Because AUROC is not differentiable, and cannot be used by non-linear optimization algorithms directly, we propose to use a suitable approximation that is differentiable. Experiments on synthetic and real data demonstrate that for the regime of interest, $n < p$, our method dominates other tested methods in all cases, and when $n > p$, it tends to smoothly transition the optimal estimate to the sample covariance matrix Σ_S, as expected, and outperforms all other methods for estimating CMs. We have also confirmed experimentally that its performance is much better than that of direct supervised classification methods, probably because the latter cannot learn very accurate decision surfaces in high-dimensional feature spaces from relatively few samples.

Although we have described the method as applicable to a specific decision problem — supervised anomaly detection where the normal data has a multivariate Gaussian distribution, but the abnormal data does not — its application is not limited to this problem only. It can also be applied to any other classification problem that matches these characteristics. Furthermore, it might be possible to extend it to more complicated distributions for the normal data, for example mixtures of multivariate Gaussians, as long as the processes of estimating the regularized CMs of the individual Gaussian components and the allocation of data points to these components can be interleaved. Another problem that is worth addressing in future research is how potentially imbalanced training sets can be handled, for example in the rather common case when very few examples of abnormal data are available.

References

[1] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, vol. 36, no. 1, pp. 199–227, 2008.

[2] J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, 1996, pp. 763–767.

[3] J. Theiler, "The incredible shrinking covariance estimator," in *Proceedings of SPIE*, vol. 839, no. 1, 2012.

[4] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, pp. 365–411, 2004.

[5] T. T. Cai, C. Zhou, and H. H. Zhou, "Optimal rates of convergence for covariance matrix estimation," *The Annals of Statistics*, vol. 38, no. 4, pp. 2118–2144, 2010.

[6] A. B. V. Chandola and V. Kumar, "Anomaly detection : a survey," *ACM Computing Surveys*, pp. 1–72, September 2009.

[7] J. Zhang, "Advancements of outlier detection: a survey," in *ICST Transactions on Scalable Information Systems*, vol. 13, no. 1, February 2013.

[8] C. L. C. Tsai, Y. Hsu and W. Lin, "Intrusion detection by machine learning: a review," *Expert Systems with Applications*, vol. 36, no. 10, pp. 11 994–12 000, 2009.

[9] R. T. T. Hastie and J. Friedman, *The Elements of Statistical Learning*. Springer, 2001.

[10] C. D. Brown and H. T. Davis, "Receiver operating characteristics curves and related decision measures: a tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 80, no. 1, pp. 24–38, 2006.

[11] M. M. L. Yan, R. Dodier and R. Wolniewicz, "Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic," in *Proceedings of the 20th International Conference on Machine Learning*, 2003.

[12] J. P. Siebert, "Vehicle recognition using rule based methods," Turing Institute Research Memorandum, Tech. Rep. TIRM-87-018, 1987.

[13] E. G. F. Kawala, A. Douzal-Chouakria and E. Dimert, "Prédictions d'activité dans les réseaux sociaux en ligne," in *4ième Conférence sur les Modèles et l'Analyse des Réseaux: Approches Mathématiques et Informatiques*, 2013, p. 16.