

# D-optimality for Minimum Volume Ellipsoid with Outliers

Alexander N. Dolia<sup>1,2</sup>, Scott F. Page<sup>1</sup>, Neil M. White<sup>1</sup>, Chris J. Harris<sup>1</sup>

<sup>1</sup>School of Electronics and Computer Science, University of Southampton,  
Southampton, SO17 1BJ, England, UK  
{ad,sfp03r,nmw,cjh}@ecs.soton.ac.uk

<sup>2</sup> Dept. 504, National Aerospace University, Kharkiv, 61070, Ukraine

## Abstract

A family of one-class classification methods is extended by the determinant maximization novelty detection (DMND) model based on the D-optimum experimental design approach for the ellipsoid estimation. Similar to the one-class classification methods based on the support vector machine or the so-called support vector data description (SVDD) approach, DMND is a method that fits a geometrical object around the training data. However, in contrast to SVDD, DMND finds the hyper-ellipsoid of the smallest volume covering the target objects that can contain outliers by maximizing the determinant of an information matrix. Simulation results are presented for the case when training data are contaminated by compactly located outliers.

## 1. Introduction

In practice, there are many applications in which it is relatively easy to collect data corresponding to target data but expensive or even economically impossible to obtain abnormal measurements that correspond to certain rare events. Therefore, the direct application of the two-class classification framework to such unbalanced datasets can lead to poor performance unless we incorporate *a priori* information about which data points belong to the target class and which are outliers.

In order to overcome these drawbacks of classical methods a number of outlier or novelty detection methods based on nonparametric [4, 2, 6], semi-parametric and parametric [6, 7] statistical approaches, support vector techniques [9, 3, 11] and neural networks [16] have been proposed.

There are two main approaches to statistical novelty detection methods based on extraction of a novelty model from the training dataset. The first approach employs information about the probability density function of features extracted from target measurements. In this case a decision about *novelty* identification is based on the probability density function (pdf)  $p(\mathbf{x}|target)$  and a pre-defined threshold  $T$  [2, 6]. A test sample is believed to be

an outlier when the level of the density there is less than  $T$ , i.e.  $p(\mathbf{x}|target) < T$ .

The performance of methods based on such an approach depends mainly on the dimensionality of data, the sample size, the choice of the novelty threshold  $T$  and how well target data is represented. Besides, usually the choice of the threshold  $T$  is based on heuristics [2]. Moreover, after learning the novelty model and setting the threshold  $T$  it is not necessary to retain knowledge of the complete pdf  $p(\mathbf{x}|target)$ , because knowledge about the support of distribution  $S_{nov}$  where  $p(\mathbf{x}|target) > T$  holds provides the same amount of information as is needed for novelty detection in the test stage. In practice, we have to carry out three steps: 1) estimate the pdf  $p(\mathbf{x}|target)$  and choose the threshold  $T$ ; 2) obtain the support of the distribution  $S_{nov}$ ; 3) apply the classification rule  $x \in S_{nov}$  to check if test point is an outlier or not.

The second approach to novelty detection is based on the last two steps, i.e. the direct estimation of the support of the distribution of the target data and analysis of whether the test point belongs to the support  $S_{nov}$  (target data) or not (outliers) [9, 3, 12]. Our approach is mainly motivated by the theory of optimal experimental design [13] and treatments of the one-class classification problem [10, 9, 11, 12], and can be considered as the extension of the minimum volume ellipsoid (MVE) algorithm based on  $D$ -optimality [13] to cover the case when data is contaminated by outliers.

The paper is organized as follows. The MVE and DMND algorithms are described in Section 2. Simulation results are demonstrated in Section 3, and conclusions are provided in Section 4.

## 2. Determinant Maximization Novelty Detection

In this section we give a short derivation of the maximum determinant novelty detection method. This is an one-class classifier inspired by the MVE method [13] and the novelty detection approach based on the support vector

machine [10, 11, 12].

There is sound theory behind the support vector machine and a number of statistical classification methods are kernelized using the so-called “kernel trick” [15, 11]. The flexibility of distribution support in the SVDD method [10, 11, 12] can be changed for example by applying different values of the smoothing parameter if a Gaussian kernel is used. When the smoothing parameter is sufficiently large the distribution support is simply a hypersphere that can be obtained by SVDD even without the kernel trick. On the other hand use of a small value for the smoothing parameter permits us to get a description of data that can be obtained by the Parzen Window approach [11]. The appropriate value of the smoothing parameter can be selected by cross-validation methods.

However, in many cases real data are not well scaled, and it is therefore difficult to expect that experimental measurements can be well described by hyperspheres. Besides, employing the kernel trick can lead to overfitting or bad generalization performance of a one-class classifier when data is indeed multivariate Gaussian contaminated by some outliers.

There are a number of classical statistical methods such as Hotelling’s  $T^2$  statistic [17], that require robust estimates of location and scatter. Please, note that strictly speaking classical Hotelling’s  $T^2$  statistic uses non-robust estimates of location and covariance matrix [1]. Therefore, it is useful to have a novelty detection method that could produce an ellipsoidal support, and and if it is necessary a more flexible description by changing just a few parameters. Herein we consider the case when the support of the training data is ellipsoidal. To describe the support of distribution we enclose the data with a  $k$ -dimensional ellipsoid with minimum volume. By minimizing the volume of hyperellipsoid we expect to reduce the rate of outlier acceptance or minimize the error of second kind.

Assume that we have a training dataset containing  $n$  samples,  $\{\mathbf{x}_i \in \mathcal{R}^{k \times 1}\}_{i=1}^n$ . In order to solve the MVE problem we need to obtain a  $(k \times k)$  positive definite matrix  $\mathbf{M} \in \mathcal{R}^{k \times k}$  and the center of the ellipsoid  $\mathbf{c}$  so as to maximize  $\det \mathbf{M}^{-1}$  subject to [13]

$$(\mathbf{x}_i - \mathbf{c})^T \mathbf{M}^{-1} (\mathbf{x}_i - \mathbf{c}) \leq k \quad (1)$$

The MVE for the dataset  $\{\mathbf{x}_i\}_{i=1}^n$  must go through at least  $k + 1$  and at most  $\frac{1}{2}k(k + 3) + 1$  support vectors. The dual optimization problem of the MVE problem (see (1)) is that of minimizing  $-\log \det \mathbf{M}(\boldsymbol{\alpha})$  with respect to  $\boldsymbol{\alpha}$ , where  $\mathbf{M}(\boldsymbol{\alpha}) = \sum_{i=1}^n \alpha_i (\mathbf{x}_i - \mathbf{c})(\mathbf{x}_i - \mathbf{c})^T$  and  $\mathbf{c} = \sum_{i=1}^n \alpha_i \mathbf{x}_i$ ;  $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$  are nonnegative numbers summing to 1 and  $\mathbf{M}(\boldsymbol{\alpha})$  might be called the “corrected” information matrix for the probability measure  $\boldsymbol{\alpha}$  (see [13] for details).

We propose to reformulate the MVE problem into

$$\begin{aligned} \min \quad & f(\mathbf{M}_d) + R^2 \\ \text{s.t.} \quad & (\mathbf{x}_i - \mathbf{c})^T \mathbf{M}_d^{-1} (\mathbf{x}_i - \mathbf{c}) \leq R^2 \end{aligned}$$

where  $\mathbf{M}_d$  is a  $k \times k$  positive definite matrix,  $R$  is an arbitrary value not equal to zero, and  $f(\mathbf{M}_d) = \log \det \mathbf{M}_d$ . Other choices of the objective function  $f(\cdot)$  can be based on the trace of the matrix  $\mathbf{M}_d$ .

In this formulation the size of the ellipsoid can be changed by varying the choice of matrices  $\mathbf{M}_d$ , the function  $f(\cdot)$  or  $R$ , but large values of  $R^2$  should be penalized. Because  $\mathbf{M}_d$  is just scaled version of the matrix  $\mathbf{M}$  we omit the index  $d$  for simplicity of notation. Then we introduce slack variables and permit some of the data samples to be outside of the ellipsoid. In order to control the volume of the ellipsoid an extra parameter  $\nu \in [0, 1]$  is introduced. In the case when  $f(\mathbf{M}_d)$  is equal to  $\log \det \mathbf{M}_d$  this leads to the following optimization problem

$$\begin{aligned} \min \quad & \varepsilon(\mathbf{M}, \mathbf{c}, R) = \log \det \mathbf{M} + R^2 + \frac{1}{\nu n} \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & (\mathbf{x}_i - \mathbf{c})^T \mathbf{M}^{-1} (\mathbf{x}_i - \mathbf{c}) \leq R^2 + \xi_i, \\ & \mathbf{M} \succ 0, \xi_i \geq 0. \end{aligned}$$

Then we use the Lagrange optimization technique in order to incorporate the constraints into  $\varepsilon(\mathbf{M}, \mathbf{c}, R)$  and construct the Lagrangian  $L$  as [11, 15]

$$\begin{aligned} L = \quad & \varepsilon(\mathbf{M}, \mathbf{c}, R) - \sum_{i=1}^n \alpha_i \{R^2 + \xi_i - \\ & (\mathbf{x}_i - \mathbf{c})^T \mathbf{M}^{-1} (\mathbf{x}_i - \mathbf{c})\} - \sum_{i=1}^n \gamma_i \xi_i, \\ & \alpha_i \geq 0, \gamma_i \geq 0 \end{aligned}$$

After setting the partial derivatives of  $L$  with respect to  $R$ ,  $\mathbf{c}$ ,  $\mathbf{M}^{-1}$  to zero and some simple algebra we obtain

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{M}^{-1}} = 0 & \implies \mathbf{M}_\alpha = \sum_{i=1}^n \alpha_i (\mathbf{x}_i - \mathbf{c})(\mathbf{x}_i - \mathbf{c})^T \\ \frac{\partial L}{\partial \mathbf{c}} = 0 & \implies \mathbf{c} = \sum_{i=1}^n \alpha_i \mathbf{x}_i \\ \frac{\partial L}{\partial R} = 0 & \implies \sum_{i=1}^n \alpha_i = 1 \\ \frac{\partial L}{\partial \xi_i} = 0 & \implies \gamma_i = \frac{1}{\nu n} - \alpha_i \\ \alpha > 0 \quad \text{and} \quad \gamma_i > 0 & \implies 0 \leq \alpha_i \leq \frac{1}{\nu n}. \end{aligned}$$

Resubstituting these values into the Lagrangian  $L$  and noting that the following factor equals to *const*

$$\sum_{i=1}^n \alpha_i (\mathbf{x}_i - \mathbf{c})^T \mathbf{M}_\alpha^{-1} (\mathbf{x}_i - \mathbf{c}) = \text{tr}(\mathbf{M}_\alpha^{-1} \mathbf{M}_\alpha) = n$$

gives us the following dual optimization problem to be satisfied by the Lagrangian multipliers  $\alpha$ :

$$\begin{aligned} \min \quad & \varepsilon(\alpha) = -\log \det \left\{ \sum_{i=1}^n \alpha_i \mathbf{x}_i \mathbf{x}_i^T - \mathbf{c} \mathbf{c}^T \right\} (2) \\ \text{s.t.} \quad & \sum_{i=1}^n \alpha_i = 1, \quad 0 \leq \alpha_i \leq \frac{1}{\nu n}, \end{aligned}$$

where  $\mathbf{c} = \sum_{i=1}^n \alpha_i \mathbf{x}_i$ . The optimization criteria  $-\log \det \mathbf{M}_\alpha$  is strictly convex on all possible nonnegative definite matrix  $\mathbf{M}_\alpha$  and therefore the optimization problem has a unique optimal solution  $\mathbf{M}_\alpha$  and  $\mathbf{c}$  but not  $\alpha$ .

If we introduce an extra dimension for the training data  $\{\mathbf{x}_i\}_{i=1}^n$  such as  $\tilde{\mathbf{x}}_i^T = [\mathbf{x}_i^T, 1]$  and  $\nu \rightarrow 0$  (no outliers) then the above optimization problem (see (2)) can be placed into  $D$ -optimum design framework [13]. Besides, when it is known *a priori* that there is no outliers ( $\nu \rightarrow 0$ ) in the dataset and data  $\{\mathbf{x}_i\}_{i=1}^n$  is centered in the origin ( $\mathbf{c} = 0$ ) in this case it is also can be considered as standard  $D$ -optimum design problem [13].

After obtaining values  $\alpha$  (see (2)) in order to check if the test point  $\mathbf{x}_t$  belongs to the estimated support we can employ the following rule

$$(\mathbf{x}_t - \mathbf{c})^T \mathbf{M}_\alpha^{-1} (\mathbf{x}_t - \mathbf{c}) \leq R^2 \quad (3)$$

where  $R^2$  is normalized by  $\mathbf{M}_\alpha$  squared distance or squared Mahalanobis distance from the center of the ellipsoid to the one of the support vectors  $\mathbf{x}_{bsv}$  that lies on its boundary:

$$R^2 = (\mathbf{x}_{bsv} - \mathbf{c})^T \mathbf{M}_\alpha^{-1} (\mathbf{x}_{bsv} - \mathbf{c}), \quad \alpha_{bsv} = \frac{1}{\nu n}.$$

### 3. Experiments

In this section we show preliminary results that characterize the performance of the proposed algorithm for target data both without and with outliers in training datasets (Fig. 1,2).

Our code was based on the **fmincon** Matlab function with a default of 400 epochs, but Newton-like and interior-point methods using linear matrix inequality constraints can be used instead [14]. The sample size of target set was  $n = 100$  (Fig.1) and there were 4 outliers (Fig.2). Compact located outliers increase the bias of location estimation. There is a clear indication that DMND method can be used to estimate a sample mean and covariance matrix when data samples are contaminated by outliers (Fig.2) if appropriate scaling is applied.

The squared Mahalanobis distance [5] is often used to find outliers in multivariate data (see Fig. 3, 4) [8]. After robust estimates of the mean and covariance matrix have been obtained the decision about whether a test point is

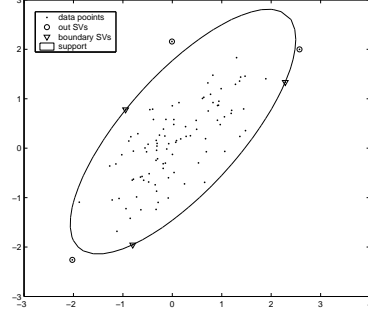


Figure 1: Illustration of the maximum determinant novelty detection approach when data are not contaminated by outliers. The ellipsoid shows the estimated distribution support and indicates which support vectors are on its boundary (*triangles*) and which fall outside (*circles*),  $n = 100, \nu = 0.05$

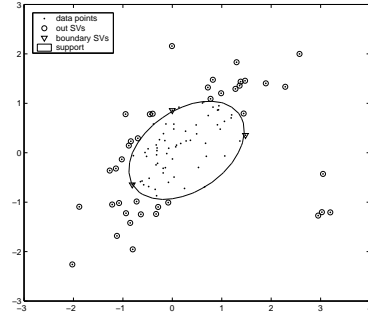


Figure 2: Illustration of the maximum determinant novelty detection approach when data are contaminated by compact outliers. The ellipsoid shows the estimated distribution support and indicates which support vectors are on its boundary (*triangles*) and which fall outside (*circles*),  $n = 104, \nu = 0.4$

an outlier or not can be based on the rule (3). It can be seen that samples that correspond to outliers have larger Mahalanobis distances (Fig. 4) than target data without outliers (Fig. 1). Therefore, squared Mahalanobis distance can be used to reject outliers.

However, our experience with the algorithm shows that the method cannot deal with a large number of outliers, and the breakdown point of method should be analyzed. It means that the DMND method does not find the MVE in all cases when training datasets are contaminated by outliers.

### 4. Conclusions

In this paper we have proposed the robust maximum determinant novelty detection method. We have presented the derivation of the robust maximum determinant novelty detection algorithm for the minimum volume ellipsoid estimation with outliers. The simulation results

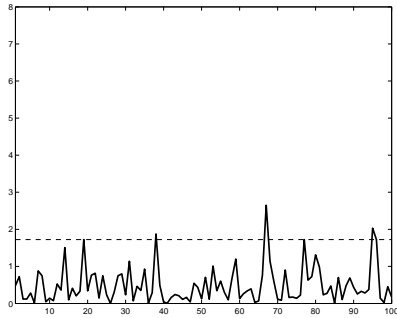


Figure 3: Squared Mahalanobis distance (see (3)) for dataset without outliers and indication of the threshold  $R^2$  (dashed line). The horizontal axis corresponds to the number of sample and the vertical axis to the value of the squared Mahalanobis distance,  $n = 100$ ,  $\nu = 0.05$

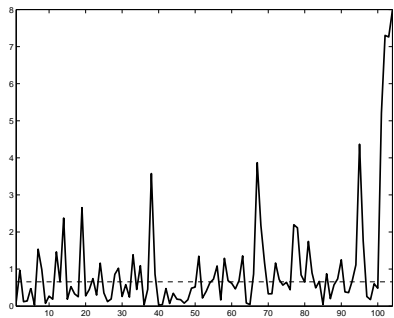


Figure 4: Squared Mahalanobis distance (see (3)) for dataset with outliers and indication of the threshold  $R^2$  (dashed line). The horizontal axis corresponds to the number of sample and the vertical axis to the value of the squared Mahalanobis distance,  $n = 104$ ,  $\nu = 0.4$

show the proposed novelty detection method based on D-optimality and slack variables can be applied to data samples contaminated by compactly located outliers.

## 5. Acknowledgements

We are grateful to many people for helpful conversations and suggestions, most notably Vladimir Vapnik, Bernhard Schölkopf, Roman Rosipal and D.M. Titterington.

## 6. References

- [1] Anderson, T. W. *An Introduction to Multivariate Statistical Analysis*, (2nd ed.) New York: Wiley (1984)
- [2] Bishop, C. Novelty Detection and Neural Network Validation, Proceedings, IEE Conference on Vision and Image Signal Processing (1994) 217–222
- [3] Campbell, C., Bennett, K.P. A Linear Programming Approach to Novelty Detection, Advances in Neural Information Processing Systems **14** MIT Press, Cambridge, MA (2001)
- [4] Devroye, L., Wise, G.L. Detection of Abnormal Behaviour via Nonparametric Estimation the Support, SIAM Journal on Applied Mathematics **38**(3) (1980) 480–488
- [5] Duda, R.O., Hart, P.E., Stork, D.G. *Pattern classification*, (2nd ed.) New York : Wiley (2001)
- [6] Nairac, A., Townsend, N., Carr, R., King, S., Cowley, P., Tarassenko L. A System for the Analysis of Jet Engine Vibration Data, Integrated Computer Aided Engineering **6** (1999) 53–65
- [7] Roberts, S.J. Novelty Detection Using Extreme Value Statistics, IEE Proceedings on Vision, Image and Signal Processing **146**(3) (1999) 124–129
- [8] Rousseeuw, P.J. and Leroy, A.M. *Robust Regression and Outlier Detection*. New York: Wiley-Interscience (1987)
- [9] Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J.S., Platt, J. Support Vector Method for Novelty Detection, In: Solla, S.A., Leen, T.K., Muller, K.R. (eds.) Neural Information Processing Systems (2000) 582–588
- [10] Schölkopf, B., Burges, C., Vapnik, V. Extracting Support Data for a Given Task, Fayyad, U.M., Uthurusamy, R. (eds.) Proceedings, First International Conference on Knowledge Discovery & Data Mining. AAAI Press, Menlo Park (1995) 252–257
- [11] Schölkopf, B., Smola, A. *Learning with Kernels*. MIT Press, Cambridge, MA (2001)
- [12] Tax, D.M.J., Duin, R.P.W. Data Domain Description by Support vectors. Verleysen, M. (ed.). Proceedings, ESANN. Brussels (1999) 251–256
- [13] Titterington, D.M. Optimal design: some geometrical aspect of D-optimality, Biometrika **62**(3) (1975) 313–320
- [14] Vandenberghe, L., Boyd, S., and Wu, S.-P. Determinant maximization with linear matrix inequality constraints, SIAM Journal on Matrix Analysis and Applications **19**(2):499-533, 1998
- [15] Vapnik, V. *Statistical Learning Theory*. Wiley NY (1998)
- [16] Ypma, A., Duin, R.P.W. Novelty Detection Using Self-organising Maps.: Progress in Connectionist Based Information Systems **2** (1998) 1322–1325
- [17] Willems, G., Pison, G., Rousseeuw, P.J., and Van Aelst, S. A Robust Hotelling Test. Metrika **55** (2002) 125–138