Domain Based LDA and QDA

Piotr Juszczak[1],
Institute for Mathematical Sciences,
Imperial College London, UK
p.juszczak@imperial.ac.uk

David M.J. Tax, Serguei Verzakov, Robert P.W. Duin
Information and Communication Theory Group,
Delft University of Technology, The Netherlands
{d.m.j.tax, s.verzakov, r.p.w.duin}@tudelft.nl

Abstract

We propose an alternative to probability density classifiers based on normal distributions LDA and QDA. Instead of estimating covariance matrices using the standard maximum likelihood estimator we estimate class domains by the minimum volume enclosing ellipsoid (ν-MVEE). The ν-MVEE is a robust statistic rejecting a specified fraction ν of the data. The performance of the domain and density approaches are compared in small sample size problems and in situations where sampling of a training and test sets is not i.i.d..

1. Introduction

The most common statistical model is the normal density [1]. According to the Central Limit Theorem, this model is correct when we assume that objects from a class originate from a single prototype and are disturbed by a large number of small independent variations. For this density model the class conditional probability $p(\mathbf{x}|\omega)$ of an object $\mathbf{x}$ given a class ω is expressed as:

$$p(\mathbf{x}|\omega) = \frac{1}{\sqrt{(2\pi)^N \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)$$

However, to estimate a mean $\boldsymbol{\mu}$ and a covariance matrix Σ accurately one must have large amount of data, especially for high dimensional problems. Moreover, the data has to be sampled i.i.d. to class distributions. As has been shown, e.g. in active learning [6], such sampling do not necessary gives the optimal performance for a fixed training set size.

Alternatively we can describe classes by their domains, which is less sensitive to the type of sampling. First, we assume that objects from a class are close in the representation space $\mathbb{R}^N$. This standard assumption is called the compactness hypothesis [2] and it characterises well behaved representations. Therefore, this suggests that we can enclose objects from a class in some kind of hull, possibly an *N*-sphere. This can be a tight description if a class is unimodal. However, we also want our model to be scale invariant. Therefore, instead of an *N*-sphere we use the affine deformations of an *N*-sphere, which is an *N*-ellipsoid.

When only class domains are estimated instead of class densities the classifiers based on class domains do not suffer from unbalanced problems and sampling of classes other than i.i.d.. However, domain based algorithm, as well as the standard estimators for the mean and covariance matrix, might heavily depend on outlier objects that are present in data. Several alternative covariance matrix estimators have been proposed in the literature, for instance the Minimum Covariance Determinant (MCD) method in [8]. This procedure is very robust and even a high fraction of outliers does not deteriorate the solution, however since based on a density it requires large amount of data sampled according to class distributions. To compute a robust class domain descriptor, which is less dependent on the type of sampling an algorithm is proposed to estimate MVEE on a fraction $1-\nu$ of data rejecting objects remote from the bulk of the data.

By replacing the standard covariance matrix with the MVEE, classifiers using the Gaussian density assumption can be redefined to new robust variations where only the Mahalanobis distance is considered.

In the next section we propose an algorithm to determine the ν-MVEE. Section 2 compares classifiers based on normal assumptions with the ones based on ν-MVEE and MCD. The performances of the classifiers in small sample size problems and problems sampled no i.i.d. are compared in section 4. The paper is concluded in section 5.

2. Robust estimation of MVEE

The minimum-volume enclosing ellipsoid problem has been studied for over 50 years. As early as 1948 (possible even earlier), [5] discussed this problem for on optimality conditions. [12] and [11] also consider the minimum volume *N*-ellipsoid problem as a special case of the more

[1]This work was done while Piotr Juszczak was at Delft University of Technology.

general maximum determinant problem. Recently MVEE has also been used in pattern recognition problems such as clustering [10]. We base the proposed ν-MVEE algorithm as the maximum determinant problem with a possibility of rejecting a specified fraction of objects.

Our concern is to cover n given points $X_t := \{\mathbf{x}_i, \mathbf{x}_i \in \mathbb{R}^N, i = 1, \ldots, n\}$ with an ellipsoid of the minimum volume for a specified fraction ν of training objects outside description. To avoid trivialities, we make the following assumption, which guarantees that a full dimensional ellipsoid can be computed in $\mathbb{R}^N$:

Assumption 1 *There is a subset of objects* $\{\mathbf{x}_1, \ldots, \mathbf{x}_{N+1}\} \subset X_t$ *which is affinely independent.*

The computation of a not fully dimensional ellipsoid is not trivial, here we focus on a full dimensional ellipsoid. The ellipsoid can be defined as:

Definition 1 *An ellipsoid* $\mathcal{E} \subseteq \mathbb{R}^N$ *is a set described by a centre* $\mathbf{c} \in \mathbb{R}^N$ *and an* $N \times N$ *symmetric positive definite matrix* E *such that*

$$\mathcal{E}_{E,\mathbf{c}} := \{\mathbf{x} \in \mathbb{R}^N | (\mathbf{x} - \mathbf{c})^T E(\mathbf{x} - \mathbf{c}) \leq 1\} \qquad (1)$$

In particular, the axes of $\mathcal{E}$ are eigenvectors of E and the length of each axis is given by $\sqrt{\lambda_i}, i = 1, \ldots, N$, where λ_i is the corresponding eigenvalue of the matrix E. We denote the positive definiteness of E by $E \succ 0$, this is equivalent to $\mathbf{x}^T E\mathbf{x} > 0,\ \forall \mathbf{x} \in \mathbb{R}^N$. When E is not positive definite the equation (1) describes any quadratic set.

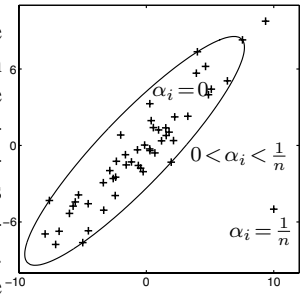


Figure 1. ν-MVEE

The volume of $\mathcal{E}_{E,\mathbf{c}}$ is given by the following formula [3]:

$$V_{\mathcal{E}_{E,\mathbf{c}}} = \frac{\pi^{\frac{N}{2}}}{\Gamma(\frac{\pi}{2}+1)} \frac{1}{\sqrt{\det(E)}} = \frac{\pi^{\frac{N}{2}}}{\Gamma(\frac{\pi}{2}+1)} \prod_{i=1}^{N} \frac{1}{\sqrt{\lambda_i}} \qquad (2)$$

where the first ratio is the volume of the unit *N*-sphere. By taking the logarithm of equation (2):

$$\ln V_{\mathcal{E}_{E,\mathbf{c}}} = \ln\left(\frac{\pi^{\frac{N}{2}}}{\Gamma(\frac{\pi}{2}+1)}\right) - \ln\sqrt{\det(E)} \qquad (3)$$

We see that minimising the volume of $\mathcal{E}$, for fixed N is equivalent to maximising the square root of the determinant of the matrix E. To get a simpler problem we change the variables:

$$M = \sqrt{E} \qquad \mathbf{z} = \sqrt{E}\mathbf{c} \qquad (4)$$

Under the assumption 1 and using (4), a natural formulation of a robust estimation of a minimum volume ellipsoid is:

$$\min_{M} \quad -\ln\det(M) + C\sum_{i=1}^{n} \xi_i, \qquad (5a)$$

$$\text{s.t.} \quad (M\mathbf{x}_i - \mathbf{z})^T(M\mathbf{x}_i - \mathbf{z}) \leq 1 + \xi_i,\ \forall_{i=1,\ldots,n}, \qquad (5b)$$

$$M \succ 0,\ \xi_i \geq 0. \qquad (5c)$$

To determine the bulk of the data and the set of potential outliers in the training set we assign a slack variable ξ_i to each objects from the training set. Additionally, a parameter C is introduced, indicating the trade-off between the volume of $\mathcal{E}$ and the sum of slacks. The value of C is crucial, it indicates whether we focus more on the minimisation of the volume of an ellipsoid $\mathcal{E}$ or on enclosing a large fraction of data. Formulation (5) can be solved using conic programming [7].

However, the optimisation (5) can be simplified further by mapping data from $\mathbb{R}^N$ to $\mathbb{R}^{N+1}$. The simplification is done by adding one additional feature, equal one, to each object $\mathbf{x}_i \in X_t$ and computing the ν-MVEE, centred at the origin. To show that these two optimisations are equivalent we denote the new set of parameters: $\mathbb{R}^N \to \mathbb{R}^{N+1},\ \mathbf{x}_i \to \tilde{\mathbf{x}}_i,\ M \to \tilde{M},\ \mathbf{z} \to \mathbf{0}$ The volume of the new ellipsoid $\mathcal{E}_{\tilde{M},\mathbf{0}}$ centred at the origin is optimised. To show that parameters of the ellipsoid $\mathcal{E}_{E,\mathbf{c}}$ can be computed from parameters of the ellipsoid $\mathcal{E}_{\tilde{M},\mathbf{0}}$ we decompose objects and the shape matrix $\tilde{M}$ as follows:

$$\tilde{\mathbf{x}}_i = \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix}, \quad \tilde{M} = \begin{pmatrix} s & \mathbf{v}^T \\ \mathbf{v} & H \end{pmatrix} \qquad (6)$$

To decide whether any object $\mathbf{x}_i$ is inside or outside ellipsoid $\mathcal{E}_{\tilde{M},\mathbf{0}}$, it is first mapped to $\mathbb{R}^{N+1}$ and then multiplied by the shape matrix $\tilde{M}$:

$$\begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix}^T \begin{pmatrix} s & \mathbf{v}^T \\ \mathbf{v} & H \end{pmatrix} \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix} = s + 2\mathbf{x}_i^T\mathbf{v} + \mathbf{x}_i^T H\mathbf{x}_i \leq 1 \qquad (7)$$

this is description of an ellipsoid in $\mathbb{R}^N$. We can rewrite (7) as:

$$(\mathbf{x}_i - \tilde{\mathbf{z}})^T \delta^{-1} H(\mathbf{x}_i - \tilde{\mathbf{z}}) \leq 1 \qquad (8)$$

where $\delta = 1 + \tilde{\mathbf{z}}^T H\tilde{\mathbf{z}} - s$ and $\tilde{\mathbf{z}} = -H^{-1}\mathbf{v}$. By comparing inequality (8) with inequality (5) we can see that:

$$\mathbf{z} = \tilde{\mathbf{z}} = -H^{-1}\mathbf{v}, \quad M = \delta^{-1}H$$

Therefore, the minimisation (5) can be written as:

$$\min_{\tilde{M},\xi_i} \quad -\ln\det(\tilde{M}) + C\sum_{i=1}^{n} \xi_i, \qquad (9a)$$

$$\text{s.t.} \quad \tilde{\mathbf{x}}_i^T \tilde{M}\tilde{\mathbf{x}}_i \leq 1 + \xi_i,\ \forall_{i=1,\ldots,n}, \qquad (9b)$$

$$\tilde{M} \succ 0,\ \xi_i \geq 0. \qquad (9c)$$

Setting the parameter C is not straightforward, there is no natural indication for its value. However, the optimisation

(9) resembles the optimisation of SVM and as in ν-SVM [9] we can modified the optimisation (9) by replacing C with the easier to set parameter ν. By using a similar trick as in [9] we modify the optimisation (9) into:

$$\min_{\tilde{M},\xi_i} \quad -\ln\det(\tilde{M}) + \frac{1}{n}\sum_{i=1}^{n}\xi_i + \nu\rho, \tag{10a}$$

$$\text{s.t.} \quad \tilde{\mathbf{x}}_i^T\tilde{M}\tilde{\mathbf{x}}_i \le \rho + \xi_i, \quad \forall_{i=1,\ldots,n}, \tag{10b}$$

$$\tilde{M} \succ 0,\ \xi_i \ge 0,\ \rho \ge 0,\ \nu \ge 0. \tag{10c}$$

where ν is now a user specified parameter that equals the fraction of objects outside the optimised ellipsoid $\mathcal{E}_{E,\mathbf{c}}$. By replacing $\tilde{M} = UU^T$, which guarantees the positiveness of $\tilde{M}$ and by assigning the derivative of the Lagrangian of (10) to zero we get $(UU^T)^{-1} = \sum_{i=1}^{n}\alpha_i\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T$. Therefore, the dual of the minimisation (10) is:

$$\max_{\alpha_i} \quad \ln\det\sum_{i=1}^{n}\alpha_i\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T, \tag{11a}$$

$$\text{s.t.} \quad \sum_{i=1}^{n}\alpha_i = \nu,\ \alpha_i \ge 0,\ \forall_i. \tag{11b}$$

Objects outside the description are determined by the optimised weights α_i. Training objects inside an ellipsoid $\mathcal{E}_{E,\mathbf{c}}$ have $\alpha_i = 0$, objects on the surface of the ellipsoid $0 < \alpha_i < \frac{1}{n}$, and objects outside the ellipsoid $\alpha_i = \frac{1}{n}$; see figure 1. The shape matrix E and the centre $\mathbf{c}$ of ellipsoid $\mathcal{E}_{E,\mathbf{c}}$ is computed from this sparse solution:

$$\tilde{M} = \begin{pmatrix} s & \mathbf{v}^T \\ \mathbf{v} & H \end{pmatrix} = \sum_{i=1}^{n}\alpha_i\mathbf{x}_i\mathbf{x}_i^T$$

$$E = (\delta^{-1}H)^T(\delta^{-1}H) \quad \text{and} \quad \mathbf{c} = -H^{-1}\mathbf{v}(E)^{-1/2}$$

Comparing the presented estimator to the existing robust solutions e.g. the minimum covariance determinant [8] the presented approach is posed as a conic problem therefore it does not need several recomputations and it gives the optimal solution for specified ν.

3 Domain based LDA and QDA

In this section we redefine normal based classifiers, LDA and QDA, using the ν-MVEE instead of a mean and a covariance matrix.

The discriminant functions $g_j(\mathbf{x})$ are computed by the linear discriminant analysis as follows [1]:

$$g_j(\mathbf{x}) = (\Sigma^{-1}\mu_j)^T\mathbf{x} - \frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j + \ln P(\omega_j) \tag{12}$$

where $P(\omega_j)$ is the prior probability. The discriminate functions $g_j(\mathbf{x})$ are computed by the quadratic discriminant analysis as follows [1]:

$$\begin{aligned} g_j(\mathbf{x}) = &-\frac{1}{2}\mathbf{x}^T\Sigma_j^{-1}\mathbf{x} + (\Sigma_j^{-1}\mu_j)^T\mathbf{x} - \frac{1}{2}\mu_j^T\Sigma_j^{-1}\mu_j \\ &-\frac{1}{2}\ln\det(\Sigma_j) + \ln P(\omega_j) \end{aligned} \tag{13}$$

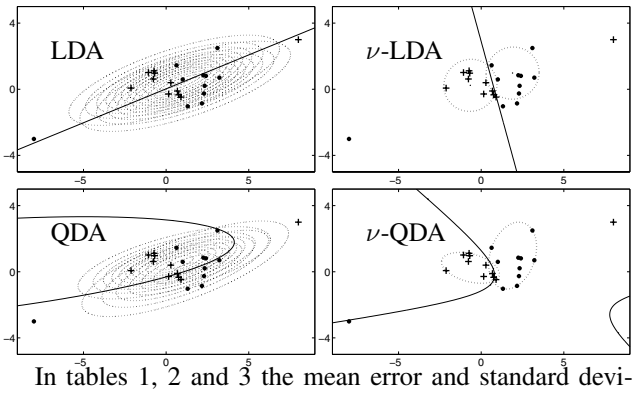
Removing all densities conditions and considering only the Mahalanobis distance, discriminant functions for the domain based QDA, which we denote ν-QDA, can be written as:

$$\begin{aligned} g_j(\mathbf{x}) &\equiv \mathbf{x}^T E_j\mathbf{x} - 2(E_j\mathbf{c}_j)^T\mathbf{x} + \mathbf{c}_j^T E_j\mathbf{c}_j, \\ \omega &\equiv \arg\min_j g_j(\mathbf{x}) \end{aligned} \tag{14}$$

where E_j and $\mathbf{c}_j$ are computed per class. $\mathbf{x}$ is assigned to the class ω with the minimum Mahalanobis distance. The ν-LDA is defined in similar way but the shape matrix E is the same for all classes and it is estimated as an element wise average of matrices E_j

4. Experiments

In this section we compare the performance of domain based classifiers ν-LDA and ν-QDA with statistically based LDA and QDA and with LDA and QDA with covariance matrices computed by the minimum covariance determinant MCD algorithm [8]. Below figures show LDA, QDA and ν-LDA, ν-QDA computed on data with a single outlier object per class. It can be seen that covariances of LDA and QDA are mainly determined by outlier objects.



In tables 1, 2 and 3 the mean error and standard deviations for domain and density based classifiers trained on several UCI repository datasets [4] are shown. In all experiments $\nu = 0.1$. In table 1, the mean error for the classifiers trained on training set with half the sizes of datasets. This represent well sampled set of problems.

In table 2, the classifiers are trained on 25% of the total size of the data. In addition we add 5% of artificial outlier objects to a training set. The outlier objects are created by moving randomly selected objects along one feature by a value of five standard deviations of that feature.

In table 3, we divide the datasets again half-half on a training and test set. However, now we enlarge the training

Table 1. Well sampled problems.

	diabetes	breast	heart	biomed	ecoli	imox	liver	satellite	waveform	letter
LDC	24.0(2.0)	4.0(0.9)	17.1(2.1)	12.4(2.0)	4.7(1.3)	10.9(2.8)	33.0(2.8)	15.0(1.8)	14.5(1.8)	30.5(1.8)
QDC	26.2(2.1)	4.5(0.6)	21.4(3.1)	10.5(2.7)	13.6(1.3)	6.0(2.0)	40.4(3.9)	14.4(2.9)	13.4(2.9)	12.4(2.9)
MCD-LDC	26.5(1.6)	3.8(1.0)	17.4(2.8)	12.7(2.9)	7.4(1.2)	10.5(1.9)	32.6(3.4)	15.6(3.4)	14.6(3.4)	32.6(3.4)
MCD-QDC	25.4(2.7)	5.1(0.9)	23.6(6.3)	18.7(4.9)	17.4(1.5)	7.0(2.0)	35.3(5.3)	14.3(5.3)	13.3(5.3)	15.3(5.3)
ν-LDC	24.4(1.9)	4.7(1.0)	17.6(2.5)	13.0(2.7)	4.4(1.7)	11.2(2.8)	32.3(3.6)	16.3(3.6)	14.3(3.6)	30.3(3.6)
ν-QDC	26.1(1.6)	4.2(1.2)	21.2(3.2)	10.9(3.2)	13.2(1.2)	6.5(3.0)	39.7(5.3)	14.7(5.3)	13.7(5.3)	12.7(5.3)

Table 2. Small sample size problems with outliers.

	diabetes	breast	heart	biomed	ecoli	imox	liver	satellite	waveform	letter
LDC	35.0(2.1)	12.0(1.4)	33.1(5.8)	22.1(2.7)	17.1(3.3)	24.9(7.3)	36.1(3.1)	34.1(2.5)	19.(2.1)	54.5(2.8)
QDC	39.2(3.1)	16.5(2.6)	35.4(4.0)	32.5(4.3)	55.3(9.3)	22.3(3.5)	43.1(2.8)	44.1(3.5)	23.1(4.1)	49.2(3.9)
MCD-LDC	29.5(2.3)	7.7(2.3)	20.2(2.7)	15.7(3.1)	15.1(2.9)	19.3(3.5)	34.2(3.5)	33.6(3.9)	17.5(3.7)	48.5(4.1)
MCD-QDC	28.4(3.1)	9.1(1.8)	28.4(5.1)	19.7(3.6)	30.4(4.5)	29.7(2.9)	38.2(5.1)	40.1(4.1)	21.4(5.4)	46.2(5.7)
ν-LDC	29.4(1.4)	7.6(1.5)	19.6(2.3)	15.0(3.3)	15.4(1.9)	13.5(3.9)	33.3(3.2)	32.4(5.1)	17.3(2.3)	46.1(3.1)
ν-QDC	28.1(2.7)	9.2(2.2)	19.2(3.8)	19.9(4.2)	21.2(2.3)	15.3(2.1)	40.7(4.3)	37.7(5.7)	20.7(5.4)	45.3(6.2)

Table 3. Problems not sampled i.i.d..

	diabetes	breast	heart	biomed	ecoli	imox	liver	satellite	waveform	letter
LDC	28.1(2.3)	7.3(1.4)	23.1(2.5)	13.4(1.1)	11.4(0.7)	15.3(1.9)	34.0(3.8)	20.1(2.9)	15.1(2.1)	32.6(7.1)
QDC	29.6(2.7)	8.3(1.3)	26.3(2.2)	14.1(3.1)	14.1(1.3)	17.1(3.1)	44.7(2.5)	16.1(1.4)	14.9(1.3)	13.4(4.1)
MCD-LDC	28.5(2.9)	7.1(1.1)	22.4(3.2)	13.6(2.2)	12.1(0.9)	15.5(2.9)	33.5(2.4)	21.4(2.9)	15.6(4.1)	31.9(4.1)
MCD-QDC	35.4(4.8)	8.1(0.5)	25.9(6.3)	14.0(2.9)	13.8(1.1)	16.0(2.2)	43.0(2.6)	15.3(2.1)	14.2(2.9)	14.1(2.1)
ν-LDC	24.9(2.1)	4.6(1.2)	17.4(2.2)	13.4(2.1)	4.9(1.8)	11.2(2.8)	32.2(1.9)	16.4(2.1)	14.3(2.4)	30.2(2.2)
ν-QDC	27.1(1.7)	4.2(1.2)	22.1(2.5)	10.8(2.7)	12.9(1.6)	9.5(3.0)	39.6(4.3)	14.6(4.1)	13.6(4.3)	12.8(3.9)

set by an additional 10% of objects from the training set i.e. 10% is doubled in the training set. The objects are selected randomly. Therefore, the density of training and test sets differ.

From table 1 it can be seen that for well sampled classification problems the performances of the classifiers were comparable. However, from results in tables 2 and 3 we can conclude that in presence of outliers, in small sample size problems and where sampling is not i.i.d. in train and test sets, the proposed domain based classifiers outperform classifiers based on density estimates.

5. Conclusions

In this paper some domain based classifiers have been proposed. The class domains are described by a minimum volume enclosing ellipsoid. As such descriptor requires estimation of only the class domain it solves much simpler problem to find a class boundary. We introduce the algorithm to compute the minimum volume ellipsoid with a specified fraction of objects outside description which allows robust statistic. It has been shown that the performance of the proposed discriminant analysis algorithm based on class domains is comparable to density approaches for well sampled problems. However, it outperforms them in situations where data is not sampled i.i.d. and in the presence of outliers.

Acknowledgements
This work was partly supported by the Dutch Organisation for Scientific Research (NWO).

References

[1] C. Bishop. *Neural networks for pattern recognition*. Oxford University Press, Walton Street, Oxford OX2 6DP, 1995.

[2] R. P. W. Duin. Compactness and complexity of pattern recognition problems. In *Internat. Symposium on Pattern Recognition*, pages 124–128, 1999.

[3] M. Grötschel, L. Lovasz, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer, 1998.

[4] S. Hettich, B. C. L., and C. J. Merz. UCI repository of machine learning databases, 1998.

[5] F. John. Extreme problems with inequalities as subsidiary conditions. *Wiley Interscience, NY*, pages 187–204, 1948.

[6] P. Juszczak and R. P. W. Duin. Selective sampling based on the variation in label assignments. In *ICPR*, pages 375–378, 2004.

[7] M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284:193–228, 1998.

[8] P. J. Rousseeuw and K. van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[9] B. Schölkopf, A. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[10] R. Shioda and L. Tuncel. Clustering via minimum volume ellipsoids. Technical report, Dep. of Combinatorics and Optimization, University of Waterloo, Canada, May 2005.

[11] K. Toh. Primal-dual path-following algorithms for determinant maximization problems with linear matrix inequalities. *Comp. Optim. App.*, 14:309–330, 1999.

[12] L. Vandenberghe, S. Boyd, and W. S. Determinant maximization with linear matrix inequality constraints. *SIAM J. Matrix Anal. App.*, 19(2):499–533, 1998.