Sorting out typicality with the inverse moment matrix SOS polynomial

Authored by:

Edouard Pauwels
Jean B. Lasserre

Abstract

We study a surprising phenomenon related to the representation of a cloud of data points using polynomials. We start with the previously unnoticed empirical observation that, given a collection (a cloud) of data points, the sublevel sets of a certain distinguished polynomial capture the shape of the cloud very accurately. This distinguished polynomial is a sum-of-squares (SOS) derived in a simple manner from the inverse of the empirical moment matrix. In fact, this SOS polynomial is directly related to orthogonal polynomials and the Christoffel function. This allows to generalize and interpret extremality properties of orthogonal polynomials and to provide a mathematical rationale for the observed phenomenon. Among diverse potential applications, we illustrate the relevance of our results on a network intrusion detection task for which we obtain performances similar to existing dedicated methods reported in the literature.

1 Paper Body

Capturing and summarizing the global shape of a cloud of points is at the heart of many data processing applications such as novelty detection, outlier detection as well as related unsupervised learning tasks such as clustering and density estimation. One of the main difficulties is to account for potentially complicated shapes in multidimensional spaces, or equivalently to account for non standard dependence relations between variables. Such relations become critical in applications, for example in fraud detection where a fraudulent action may be the dishonest combination of several actions, each of them being reasonable when considered on their own. Accounting for complicated shapes is also related to computational geometry and nonlinear algebra applications, for example integral computation [11] and reconstruction of sets from moments data [6, 7, 12]. Some of these problems have connections and potential applications in machine learning. The work presented in this paper brings together ideas from both disciplines, leading to a method which allows to encode in a simple manner the global shape and spatial concentration of points within a cloud. We start

with a surprising (and apparently unnoticed) empirical observation. Given a collection of points, one may build up a distinguished sum-of-squares (SOS) polynomial whose coefficients (or Gram matrix) is the inverse of the empirical moment matrix (see Section 3). Its degree depends on how many moments are considered, a choice left to the user. Remarkably its sublevel sets capture much of the global shape of the cloud as illustrated in Figure 3. This phenomenon is not incidental as illustrated in many additional examples in Appendix A. To the best of our knowledge, this observation has remained unnoticed and the purpose of this paper is to report this empirical finding to the machine learning community and provide first elements toward a mathematical understanding as well as potential machine learning applications. 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

146370 67380
30670
30 67 67 38 0 0 14 63 70
35
13810 6860
3570
1850
950 460
340
210
333
80 10
13
70 18 50 95 81 0 0 68 60
333340
30 15
138
10
357 0 6860
185
0
950

Figure 1: Left: 1000 points in R2 and the levelsets of the corresponding inverse moment matrix SOS polynomial Q?,d (d = 4). The level set p+d d , which corresponds to the average value of Q?,d , is represented in red. Right: 1040 points in R2 with size and color proportional to the value of inverse moment matrix SOS polynomial Q?,d (d = 8). The proposed method is based on the computation of the coefficients of a very specific polynomial which depends solely on the empirical moments associated with the data points. From a practical perspective, this can be done via a single pass through the data, or even in an online fashion via a sequence of efficient Woodbury updates. Furthermore the computational cost of evaluating the polynomial does not depend on the number of data points which is a crucial difference with existing non-parametric methods such as nearest neighbors or kernel based methods [3]. On

the other hand, this computation requires the inversion of a matrix whose size depends on the dimension of the problem (see Section 3). Therefore, the proposed framework is suited for moderate dimensions and potentially very large number of observations. In Section 4 we first describe an affine invariance result which suggests that the distinguished SOS polynomial captures very intrinsic properties of clouds of points. In a second step, we provide a mathematical interpretation that supports our empirical findings based on connections with orthogonal polynomials [5]. We propose a generalization of a well known extremality result for orthogonal univariate polynomials on the real line (or the complex plane) [16, Theorem 3.1.2]. As a consequence, the distinguished SOS polynomial of interest in this paper is understood as the unique optimal solution of a convex optimization problem: minimizing an average value over a structured set of positive polynomials. In addition, we revisit [16, Theorem 3.5.6] about the Christoffel function. The mathematics behind provide a simple and intuitive explanation for the phenomenon that we empirically observed. Finally, in Section 5 we perform numerical experiments on KDD cup network intrusion dataset [13]. Evaluation of the distinguished SOS polynomial provides a score that we use as a measure of outlyingness to detect network intrusions (assuming that they correspond to outlier observations). We refer the reader to [3] for a discussion of available methods for this task. For the sake of a fair comparison we have reproduced the experiments performed in [18] for the same dataset. We report results similar to (and sometimes better than) those described in [18] which suggests that the method is comparable to other dedicated approaches for network intrusion detection, including robust estimation and Mahalanobis distance [8, 10], mixture models [14] and recurrent neural networks [18].

2

Multivariate polynomials, moments and sums of squares

Notations: We fix the ambient dimension to be p throughout the text. For example, we will manipulate vectors in Rp as well as p-variate polynomials with real coefficients. We denote by X a set of p variables X1 , . . . , Xp which we will use in mathematical expressions defining polynomials. We identify monomials from the canonical basis of p-variate polynomials with their exponents in ?p ?1 ?2 p ? Np : we associate Pp to ? = (?i)i=1...p ? N the monomial X := X1 X2 . . . Xp which degree is deg(?) := i=1 ?i . We use the expressions ¡gl and ?gl to denote the graded lexicographic order, a well ordering over p-variate monomials. This amounts to, first, use the canonical order on the 2

degree and, second, break ties in monomials with the same degree using the lexicographic order with X1 = a, X2 = b . . . For example, the monomials in two variables X1 , X2 , of degree less or equal to 3 listed in this order are given by: 1, X1 , X2 , X12 , X1 X2 , X22 , X13 , X12 X2 , X1 X22 , X23 . We denote by Npd , the set {? ? Np ; deg(?) ? d} ordered by ?gl . R[X] denotes the set of p-variate polynomials: linear combinations of monomials with real coefficients. The degree of a polynomial is the highest of the degrees of its monomials with nonzero coefficients1 . We use the same notation, deg(?), to denote the degree of a polynomial or of an element of Np . For d ? N, Rd [X] denotes the set of p-variate polynomials of degree less or equal to d. We set s(d) = p+d d , the

number of monomials of degree less or equal to d. We will denote by vd (X) the vector of monomials of degree less or equal to d sorted by ?gl . We let vd (X) := (X ?)??Np ? Rd [X]s(d) . With this notation, d we can write a polynomial P ? Rd [X] as follows P (X) = hp, vd (X)i for some real vector of coefficients p = (p?)??Np ? Rs(d) ordered using ?gl . Given x = (xi)i=1...p ? Rp , P (x) denotes d the evaluation of P with the assignments X1 = x1 , X2 = x2 , . . . Xp =R xp . Given a Borel probability measure ? and ? ? Np , y? (?) denotes the moment ? of ?: y? (?) = Rp x? d?(x). Throughout the paper, we will only consider measures of which all moments are finite. Moment matrix: Given a Borel probability measure ? on Rp , the moment matrix of ?, Md (?), is a matrix indexed by monomials of degree at most d ordered by ?gl . For ?, ? ? Npd , the corresponding entry in Md (?) is defined by Md (?)?,? := y?+? (?), the moment ? + ? of ?. When p = 2, letting y? = y? (?) for ? ? N24 , we have

M2 (?) :
1 X1 X2 X12 X1 X2 X22
1
X1
X2
X12
X1 X2
1 y10 y01 y20 y11 y02
y10 y20 y11 y30 y21 y12
y01 y11 y02 y21 y12 y03
y20 y30 y21 y40 y31 y22
y11 y21 y12 y31 y22 y13
X22 y02 y12 y03 . y22 y13 y04

Md (?) is positive semidefinite for all d ? N. Indeed, for any Rp ? Rs(d) , let P ? Rd [X] be the polynomial with vector of coefficients p, we have pT Md (?)p = Rp P 2 (x)d?(x) ? 0. Furthermore, R we have the identity Md (?) = Rp vd (x)vd (x)T d?(x) where the integral is understood elementwise. Sum of squares (SOS): We denote by ?[X] ? R[X] (resp. ?d [X] ? Rd [X]), the set of polynomials (resp. polynomials of degree at most d) which can be written as a sum of squares of polynomials. Let P ? R2m [X] for some m ? N, then P belongs to ?2m [X] P if there exists a finite J ? N and a family of polynomials Pj ? Rm [X], j ? J, such that P = j?J Pj2 . It is obvious that sum of squares polynomials are always nonnegative. A further interesting property is that this class of polynomials is connected with positive semidefiniteness. Indeed, P belongs to ?2m [X] if and only if ?Q ? Rs(m)?s(m) , Q 0, P (x) = vd (x)T Qvd (x), ?x ? Rp . As a consequence, every positive semidefinite matrix Q ? R ?2m [X] by using the representation in (1).

3
s(m)?s(m)
(1)
defines a polynomial in
Empirical observations on the inverse moment matrix SOS polynomial

The inverse moment-matrix SOS polynomial is associated to a measure ? which satisfies the following. Assumption 1 ? is a Borel probability measure on Rp with all its moments finite and Md (?) is positive definite for a given d ? N. Definition 1 Let ?, d satisfy Assumption 1. We call the SOS polynomial Q?,d ? ?2d [X] defined by the application: x 7?

Q?,d (x) := vd (x)T Md (?)?1 vd (x),

1

x ? Rp ,

(2)

For the null polynomial, we use the convention that its degree is 0 and it is ?gl smaller than all other monomials.

3

the inverse moment-matrix SOS polynomial of degree 2d associated to ?. Actually, connection to orthogonal polynomials will show that the inverse function x 7? Q?,d (x)?1 is called the Christoffel function in the literature [16, 5] (see also Section 4). In the remainder of this section, we focus on the situation when ? corresponds to an empirical measureP over n points in Rp which are fixed. So let x1 , . . . , xn ? Rp be a fixed set of points and let n 1 ? := n i=1 ?xi where ?x corresponds to the Dirac measure at x. In such a case the polynomial Q?,d in (2) is determined only by the empirical moments up to degree 2d of our collection of points. Note that we also require that Md (?) 0. In other words, the points x1 , . . . , xn do not belong to an algebraic set defined by a polynomial of degree less or equal to d. We first describe empirical properties of inverse moment matrix SOS polynomial in this context of empirical measures. A mathematical intuition and further properties behind these observations are developped in Section 4. 3.1

Sublevel sets

The starting point of our investigations is the following phenomenon which to the best of our knowledge has remained unnoticed in the literature. For the sake of clarity and simplicity we provide an illustration in the plane. Consider the following experiment in R2 for a fixed d ? N: represent on the same graphic, the cloud of points {xi }i=1...n and the sublevel sets of SOS polynomial Q?,d in R2 (equivalently, the superlevel sets of the Christoffel function). This is illustrated in the left panel of Figure 3. The collection of points consists of 500 simulations of two different Gaussians and the value of d is 4. The striking feature of this plot is that the level sets capture the global shape of the cloud of points quite accurately. In particular, the level set {x : Q?,d (x) ? p+d d } captures most of the points. We could reproduce very similar observations on different shapes with various number of points in R2 and degree d (see Appendix A). 3.2

Measuring outlyingness

An additional remark in a similar line is that Q?,d tends to take higher values on points which are isolated from other points. Indeed in the left panel of Figure 3, the value of the polynomial tends to be smaller on the boundary of the cloud. This extends to situations where the collection of points correspond to shape with a high density of points with a few additional outliers. We reproduce a similar experiment on the right panel of Figure 3. In this example, 1000

points are sampled close to a ring shape and 40 additional points are sampled uniformly on a larger square. We do not represent the sublevel sets of Q?,d here. Instead, the color and shape of the points are taken proportionally to the value of Q?,d , with d = 8. First, the results confirm the observation of the previous paragraph, points that fall close to the ring shape tend to be smaller and points on the boundary of the ring shape are larger. Second, there is a clear increase in the size of the points that are relatively far away from the ring shape. This highlight the fact that Q?,d tends to take higher value in less populated areas of the space. 3.3

Relation to maximum likelihood estimation

If we fix d = 1, we recover the maximum Pn likelihood estimation Pn for the Gaussian, up to a constant additive factor. To see this, set ? = n1 i=1 xi and S = n1 i=1 xi xTi . With this notation, we have the following block representation of the moment matrix,

1 + ?T V ?1 ? ??T V ?1 1 ?T ?1 Md (?) = Md (?) = , ? S ?V ?1 ? V ?1 where V = S ? ??T is the empirical covariance matrix and the expression for the inverse is given by Schur complement. In this case, we have Q?,1 (x) = 1 + (x ? ?)T V ?1 (x ? ?) for all x ? Rp . We recognize the quadratic form that appears in the density function of the multivariate Gaussian with parameters estimated by maximum likelihood. This suggests a connection between the inverse SOS moment polynomial and maximum likelihood estimation. Unfortunately, this connection is difficult to generalize for higher values of d and we do not pursue the idea of interpreting the empirical observations of this section through the prism of maximum likelihood estimation and leave it for further research. Instead, we propose an alternative view in Section 4. 4

3.4

Computational aspects

Recall that s(d) = p+d is the number of p-variate monomials of degree up to d. The computation d of Q?,d requires O(ns(d)2) operations for the computation of the moment matrix and O(s(d)3) operations for the matrix inversion. The evaluation of Q?,d requires O(s(d)2) operations. Estimating the coefficients of Q?,d has a computational cost that depends only linearly in the number of points n. The cost of evaluating Q?,d is constant with respect to the number of points n. This is an important contrast with kernel based or distance based methods (such as nearest neighbors and one class SVM) for density estimation or outlier detection since they usually require at least O(n2) operations for the evaluation of the model [3]. Moreover, this is well suited for online settings where inverse moment matrix computation can be done using rank one Woodbury updates [15, Section 2.7.1]. The dependence in the dimension p is of the order of pd for a fixed d. Similarly, the dependence in d is of the order of dp for a fixed dimension p and the joint dependence is exponential. Furthermore, Md (?) has a Hankel structure which is known to produce ill conditioned matrices. This suggests that the direct computation and evaluation of Q?,d will mostly make sense for moderate dimensions and degree d. In our experiments, for large d, the evaluation of Q?,d remains quite stable, but the inversion leads to numerical error for higher values (around 20).

4

Invariance and interpretation through orthogonal polynomials

The purpose of this section is to provide a mathematical rationale that explains the empirical observations made in Section 3. All the proofs are postponed to Appendix B. We fix a Borel probability measure ? on Rp which satisfies Assumption 1. Note that Md (?) is always positive definite if ? is not supported on the zero set of a polynomial of degree at most d. Under Assumption 1, Md (?) induces an inner product on Rs(d) and by extension on Rd [X] (see Section 2). This inner product is denoted by h?, ?i? and satisfies for any polynomials P, Q ? Rd [X] with coefficients p, q ? Rs(d) , Z hP, Qi? := hp, Md (?)qiRs(d) = P (x)Q(x)d?(x). Rp

We will also use the canonical inner product over Rd [X] which we write hP, QiRd [X] := hp, qiRs(d) for any polynomials P, Q ? Rd [X] with coefficients p, q ? Rs(d) . We will omit the subscripts for this canonical inner product and use h?, ?i for both products. 4.1

Affine invariance

It is worth noticing that the mapping x 7? Q?,d (x) does not depend on the particular choice of vd (X) as a basis of Rd [X], any other basis would lead to the same mapping. This leads to the result that Q?,d captures affine invariant properties of ?. Lemma 1 Let ? satisfy Assumption 1 and A ? Rp?p , b ? Rp define an invertible affine mapping on Rp , A : x ? Ax+b. Then, the push foward measure, defined by ? ?(S) = ?(A?1 (S)) for all Borel sets S ? Rp , satisfies Assumption 1 (with the same d as ?) and for all x ? Rp , Q?,d (x) = Q??,d (Ax + b). Pn Lemma 1 is probably better understood when ? = 1/n i=1 ?xi as in Section 3. In this case, we Pn have ? ? = 1/n i=1 ?Axi +b and Lemma 1 asserts that the level sets of Q??,d are simply the images of those of Q?,d under the affine transformation x 7? Ax + b. This is illustrated in Appendix D. 4.2

Connection with orthogonal polynomials

We define a classical [16, 5] family of orthonormal polynomials, {P? }??Np ordered according to ?gl d which satisfies for all ? ? Npd hP? , X ? i = 0 if ? ¡gl ?, hP? , P? i? = 1, hP? , X ? i? = 0 if ? ¡gl ?, hP? , X ? i? ¿ 0.

(3)

It follows from (3) that hP? , P? i? = 0 if ? 6= ?. Existence and uniqueness of such a family is guaranteed by the Gram-Schmidt orthonormalization process following the ?gl order, and by the 5

positivity of the moment matrix, see for instance [5, Theorem 3.1.11]. There exist determinantal formulae [9] and more precise description can be made for measures which have additional geometric properties, see [5] for many examples. Let Dd (?) be the lower triangular matrix whose rows are the coefficients of the polynomials P? defined in (3) ordered by ?gl . It can be shown that Dd (?) = Ld (?)?T , where Ld (?) is the Cholesky factorization of Md (?). Furthermore, there is a direct relation with the inverse moment matrix as Md (?)?1 = Dd (?)T Dd (?) [9, Proof of Theorem 3.1]. This has the following consequence. P Lemma 2 Let ? satisfy Assumption 1, then Q?,d = P?2 , where the family {P? }??Np is ??Np d d R defined by (3) and Rp Q?,d (x)d?(x) = s(d). That is, Q?,d

is a very specific and distinguished SOS polynomial, the sum of squares of the orthonormal basis elements {P? }??Np of Rd (X) (w.r.t. ?). Furthermore, the average value of Q?,d d with respect to ? is s(d) which corresponds to the red level set in left panel of Figure 3. 4.3

A variational formulation for the inverse moment matrix SOS polynomial

In this section, we show that the family of polynomials {P? }??Np defined in (3) is the unique d solution (up to a multiplicative constant) of a convex optimization problem over polynomials. This fact combined with Lemma 2 provides a mathematical rationale for the empirical observations outlined in Section 3. Consider the following optimization problem. Z X 1 min p Q? (x)2 d?(x) (4) Q? ,?? ,??Nd 2 Rp p ??Nd X s.t. q?? ? exp(??), q?? = 0, ?, ? ? Npd , ? ¡gl ?, ?? = 0, ??Np d

P where Q? (x) = ??Np q?? x? is a polynomial and ?? is a real variable for each ? ? Npd . We first d P comment on problem (4). Let P = ??Np Q2? be the SOS polynomial appearing in the objective d function of (4). The objective of (4) simply involves the average value of P with respect to ?. Let Sd ? ?d [X] be the set of such SOS polynomials P which have a sum of square decomposition satisfying the constraints of (4) (for some arbitrary value of the real variables {?? }??Np). With this d R notation, problem (4) has the simple formulation minP ?Sd 21 P d?. Based on this formulation, problem (4) can be interpreted as balancing two antagonist targets. On one hand the minimization of the average value of the SOS polynomial P with respect to ?, on the other hand the avoidance of the trivial polynomial, enforced by the constraint that P ? Sd . The constraint P ? Sd is simple and natural. It ensures that P is a sum of squares of polynomials {Q? }??Np , where d the leading term of each Q? (according to the ordering ?gl) is q?? x? with q?? ¿ 0 (and hence does not vanish). Inversely, using Cholesky factorization, for any SOS polynomial Q of degree 2d which coefficient matrix (see equation (1)) is positive definite, there exists a ¿ 0 such that aQ ? Sd . This suggests that Sd is a quite general class of nonvanishing SOS polynomials. The following result, which gives a relation between Q?,d and solutions of (4), uses a generalization of [16, Theorem 3.1.2] to several orthogonal polynomials of several variables. Theorem 1 : Under Assumption 1, problem (4) is?a convex optimization problem with a unique ? optimal solution (Q?? , ???), which satisfies Q Npd , for some ? ¿ 0. In particular, ?, ? ? P? = ?P P the distinguished SOS polynomial Q?,d = P?2 = ?1 ??Np (Q??)2 , is (part of) the unique ??Np d d optimal solution of (4). Theorem 1 states that up to the scaling factor ?, the distinguished SOS polynomial Q?,d is the unique optimal solution of problem (4). A detailed proof is provided in the Appendix B and we only sketch the main ideas here. First, it is remarkable that for each fixed ? ? Npd (and again up ? is the uniqueo optimal solution of the problem: n Rto a scaling factor) the polynomial PP 2 ? minQ Q d? : Q ? Rd [X], Q(x) = x + ?¡gl ? q? x? . This fact is well-known in the univariate case [16, Theorem 3.1.2] and does not seem to have been exploited in the literature, at 6

least for purposes similar to ours. So intuitively, P?2 should be as close to 0 as possible on the support of ?. Problem (4) has similar properties and the R

constraint on the vector of weights ? enforces that, at an optimal solution, the contribution of (Q??)2 d? to the overall sum in the criterion is the same for all ?. Using Lemma 2 yields (up to a multiplicative constant) the polynomial Q?,d . Other constraints on ? would yield different weighted sum of the squares P?2 . This will be a subject of further investigations. To sum up, Theorem 1 provides a rationale for our observations. Indeed when solving (4), intuitively, Q?,d should be close to 0 on average while remaining in a class of nonvanishing SOS polynomials. 4.4

Christoffel function and outlier detection

The following result from [5, Theorem 3.5.6] draws a direct connection between Q?,d and the Chritoffel function (the right hand side of (5)). Theorem 2 ([5]) Let Assumption 1 hold and let z ? Rp be fixed, arbitrary. Then Z

?1 2 Q?,d (z) = min P (x) d?(x) : P (z) = 1 . P ?Rd [X]

(5)

Rp

Theorem 2 provides a mathematical rationale for the use of Q?,d for outlier or novelty detection purposes. Indeed, from Lemma 2 and equation (3), we have Q?,d ? 1 on Rp . Furthermore, the solution of the minimization problem in (5) satisfies P (z)2 = 1 and ? x ? Rp : P (x)2 ? 1 ? inequality). Hence, for high values of Q?,d (z), the sublevel set 1 ? Q?,d (z)?1 (by Markov?s

x ? Rp : P (x)2 ? 1 contains most of the mass of ? while P (z)2 = 1. An illustration of this discussion is given in appendix E. Again the result of Theorem 2 does not seem to have been interpreted for purposes similar to ours.

5

Experiments on network intrusion datasets

In addition to having its own mathematical interest, Theorem 1 can be exploited for various purposes.

p+d p For instance, the sub-level sets of Q?,d , and in particular {x ? R : Q?,d (x) ? d }, can be used to encode a cloud of points in a simple and compact form. However in this section we focus on another potential application in anomaly detection. Empirical findings described in Section 3 suggest that the polynomial Q?,d can be used to detect outliers in a collection of real vectors (with ? the empirical average). This is backed up by the results presented in Section 4. We illustrate these properties on a real world example. We choose the KDD cup 99 network intrusion dataset [13] consisting of network connection data, labeled as normal traffic or network intrusions. We follow [19] and [18] and construct five datasets consisting of labeled vectors in R3 with the following properties Dataset Number of examples Proportions of attacks

http 567498 0.004

smtp 95156 0.0003

ftp-data 30464 0.023

ftp 4091 0.077

others 5858 0.016

The details on the datasets construction are available in [19, 18] and reproduced in Appendix C. The main idea is to compute an outlyingness score

(independant of the label) and compare outliers predicted by the score and network intrusion labels. The underlying assumption is that network intrusions correspond to infrequent abnormal behaviors and could be considered as outliers. We reproduce the same experiment as in [18, Section 5.4] using the value of Q?,d from Definition 1 as an outlyingness score (with d = 3). The authors of [18] have compared different methods in the same experimental setting: robust estimation and Mahalanobis distance [8, 10], mixture models [14] and recurrent neural networks. The results are gathered in [18, Figure 7]. In the left panel of Figure 2 we represent the same performance measure for our approach: we first compute the value of Q?,d for each datapoint and use it as an outlyingness score. We then display the proportion of correctly identified outliers, with score above a given threshold, as a function of the proportion of examples with score above the threshold (for different values of the threshold). The main comments are as follows. 7

d (AUPR)
dataset
0.8
http 0.6
smtp ftp_data
0.4
ftp
2 (0.18) 3 (0.18)
0.50
4 (0.16) 5 (0.15)
0.25
others
0.2
1 (0.08)
0.75
Precision
% correctly identified outliers
1.00
1.0
6 (0.13)
0.0
0.00 0.0
0.2
0.4
0.6
0.8
1.0
0.0
% top outlyingness score
0.2
0.4
0.6

0.8
1.0
Recall

Figure 2: Left: reproduction of the results described in [18] with the evaluation of Q?,d as an outlyingness score (d = 3). Right: precision-recall curves for different values of d (dataset ?others?). ? The inverse moment matrix SOS polynomial does detect network intrusions with varying performances on the five datasets. ? Except for the ?ftp-data dataset?, the global shape of these curves are very similar to results reported in [18, Figure 7] indicating that the proposed approach is comparable to other dedicated methods for intrusion detection in these four datasets. In a second experiment, we investigate the effect of changing the value of d on the performances. We focus on the ?others? dataset because it is the most heterogeneous. We adopt a slightly different measure of performance and use precision recall (see for example [4]) to measure performances in identifying network intrusions (the higher the curve, the better). We call the area under such curves the AUPR. The right panel of Figure 2 represents these results. First, the case d = 1, which corresponds to vanilla Mahalanobis distance as outlined in Section 3.3, gives poor performances. Second, the global performances rapidly increase with d and then decrease and stabilize. This suggests that d can be used as a tuning parameter to control the ?complexity? of Q?,d . Indeed, 2d is the degree of the polynomial Q?,d and it is expected that more complex models will identify more diverse classes of examples as outliers. In our case, this means identifying regular traffic as outliers while it actually does not correspond to intrusions. In general, a good heuristic regarding the tuning of d is to investigate performances on a well specified task in a preliminary experiment.

6

Future work

An important question is the asymptotic regime when d ? ?. Current state of knowledge suggests that, up to a correct scaling, the limit of the Christoffel functions (when known to exist) involves an edge effect term, related to the support of the measure, and the density of ? with respect to Lebesgue measure, see for example [2] for the Euclidean ball. It also suggests connections with the notion of equilibrium measure in potential theory [17, 1, 7]. Generalization and interpretation of these results in our context will be investigated in future work. Even though good approximations are obtained with low degree (at least in dimension 2 or 3), the approach involves the inversion of large ill conditioned Hankel matrices which reduces considerably the applicability for higher degrees and dimensions. A promising research line is to develop approximation procedures and advanced optimization and algebra tools so that the approach could scale computationally to higher dimensions and degrees. Finally, we did not touch the question of statistical accuracy. In the context of empirical processes, this will be very relevant to understand further potential applications in machine learning and reduce the gap between the abstract orthogonal polynomial theory and practical machine learning applications. Acknowledgments This work was partly supported by project ERC-ADG TAMING 666981, ERC-Advanced

Grant of the European Research Council and grant number FA9550-15-1-0500 from the Air Force Office of Scientific Research, Air Force Material Command.
8

2 References

[1] R. J. Berman (2009). Bergman kernels for weighted polynomials and weighted equilibrium measures of Cn . Indiana University Mathematics Journal, 58(4):1921?1946. [2] L. Bos, B. Della Vecchia and G. Mastroianni (1998). On the asymptotics of Christoffel functions for centrally symmetric weights functions on the ball in Rn . Rendiconti del Circolo Matematico di Palermo, 52:277?290. [3] V. Chandola, A. Banerjee and V. Kumar (2009). Anomaly detection: A survey. ACM computing surveys (CSUR) 41(3):15. [4] J. Davis and M. Goadrich (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning (pp. 233-240). ACM. [5] C.F. Dunkl and Y. Xu (2001). Orthogonal polynomials of several variables. Cambridge University Press. MR1827871. [6] G.H Golub, P. Milanfar and J. Varah (1999). A stable numerical method for inverting shape from moments. SIAM Journal on Scientific Computating 21(4):1222?1243 (1999). [7] B. Gustafsson, M. Putinar, E. Saff and N. Stylianopoulos (2009). Bergman polynomials on an archipelago: estimates, zeros and shape reconstruction. Advances in Mathematics 222(4):1405? 1460. [8] A.S. Hadi (1994). A modification of a method for the detection of outliers in multivariate samples. Journal of the Royal Statistical Society. Series B (Methodological), 56(2):393-396. [9] J.W. Helton, J.B. Lasserre and M. Putinar (2008). Measures with zeros in the inverse of their moment matrix. The Annals of Probability, 36(4):1453-1471. [10] E.M. Knorr, R.T. Ng and R.H.Zamar (2001). Robust space transformations for distance-based operations. Proceedings of the international conference on Knowledge discovery and data mining (pp. 126-135). ACM. [11] J.B. Lasserre (2015). Level Sets and NonGaussian Integrals of Positively Homogeneous Functions. International Game Theory Review, 17(01):1540001. [12] J.B. Lasserre and M.Putinar (2015). Algebraic-exponential Data Recovery from Moments. Discrete & Computational Geometry, 54(4):993-1012. [13] M. Lichman (2013). UCI Machine Learning Repository, http://archive.ics.uci.edu/ml University of California, Irvine, School of Information and Computer Sciences. [14] J.J. Oliver, R.A.Baxter and C.S. Wallace (1996). Unsupervised learning using MML. Proceedings of the International Conference on Machine Learning (pp. 364-372). [15] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery (2007). Numerical Recipes: The Art of Scientific. Computing (3rd Edition). Cambridge University Press. [16] G. Szeg? (1974). Orthogonal polynomials. In Colloquium publications, AMS, (23), fourth edition. [17] V. Totik (2000). Asymptotics for Christoffel functions for general measures on the real line. Journal d?Analyse Math?matique, 81(1):283-303. [18] G. Williams, R. Baxter, H. He, S. Hawkins and L. Gu (2002). A Comparative Study of RNN for Outlier Detection in Data Mining. IEEE International Conference on Data Mining (p. 709). IEEE

Computer Society. [19] K. Yamanishi, J.I. Takeuchi, G. Williams and P. Milne (2004). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Mining and Knowledge Discovery, 8(3):275-300.

9