Sorting out typicality with the inverse moment matrix SOS polynomial

Jean-Bernard Lasserre, Edouard Pauwels

Sorting out typicality with the inverse moment matrix SOS polynomial

Jean-Bernard Lasserre
LAAS-CNRS & IMT
Université de Toulouse
31400 Toulouse, France
lasserre@laas.fr

Edouard Pauwels
IRIT & IMT
Université Toulouse 3 Paul Sabatier
31400 Toulouse, France
edouard.pauwels@irit.fr

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 Introduction

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 [9] and reconstruction of sets from moments data [4, 5, 10]. 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.

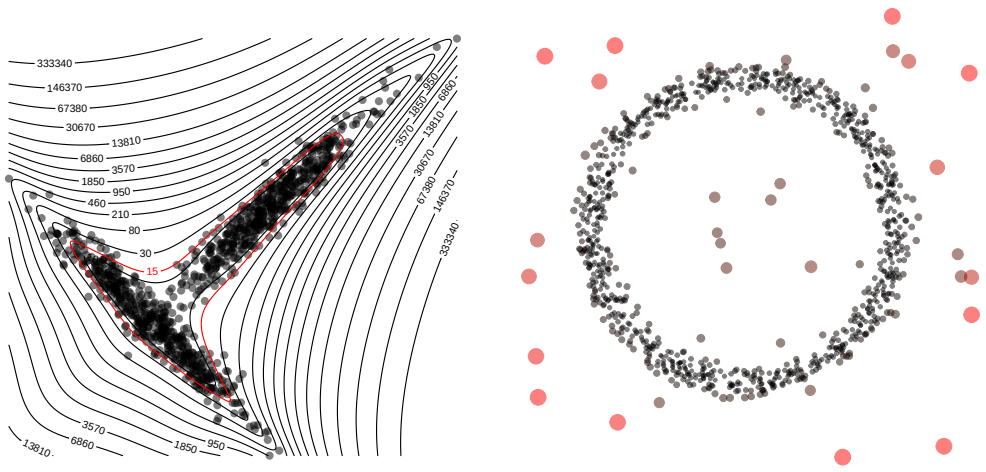

Figure 1: Left: 1000 points in $\mathbb{R}^2$ and the level sets of the corresponding inverse moment matrix SOS polynomial $Q_{\mu,d}$ ($d = 4$). The level set $\binom{p+d}{d}$, which corresponds to the average value of $Q_{\mu,d}$, is represented in red. Right: 1040 points in $\mathbb{R}^2$ with size and color proportional to the value of inverse moment matrix SOS polynomial $Q_{\mu,d}$ ($d = 8$).

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 nonparametric methods such as nearest neighbors or kernel based methods [1]. 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 [3]. We propose a generalization of a well known extremality result for orthogonal univariate polynomials on the real line (or the complex plane) [13, 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 [13, 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 [11]. 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 [1] for a discussion of available methods for this task. For the sake of a fair comparison we have reproduced the experiments performed in [14] for the same dataset. We report results similar to (and sometimes better than) those described in [14] which suggests that the method is comparable to other dedicated approaches for network intrusion detection, including robust estimation and Mahalanobis distance [6, 8], mixture models [12] and recurrent neural networks [14].

2 Multivariate polynomials and moments

2.1 Notations

We fix the ambient dimension to be p throughout the text. For example, we will manipulate vectors in $\mathbb{R}^p$ as well as p-variate polynomials with real coefficients. We denote by X a set of p variables $X_1, \ldots, X_p$ which we will use in mathematical expressions defining polynomials. We identify

monomials from the canonical basis of p-variate polynomials with their exponents in $\mathbb{N}^p$: we associate to $\alpha = (\alpha_i)_{i=1\dots p} \in \mathbb{N}^p$ the monomial $X^\alpha := X_1^{\alpha_1} X_2^{\alpha_2} \dots X_p^{\alpha_p}$ which degree is $\deg(\alpha) := \sum_{i=1}^p \alpha_i$. We use the expressions $<_{gl}$ and $\leq_{gl}$ to denote the graded lexicographic order, a well ordering over p-variate monomials. This amounts to, first, use the canonical order on the degree and, second, break ties in monomials with the same degree using the lexicographic order with $X_1 = a, X_2 = b \dots$ For example, the monomials in two variables X_1, X_2, of degree less or equal to 3 listed in this order are given by: 1, X_1, X_2, X_1^2, X_1X_2, X_2^2, X_1^3, $X_1^2X_2$, $X_1X_2^2$, X_2^3.

We denote by $\mathbb{N}^p_d$, the set $\{\alpha \in \mathbb{N}^p;\ \deg(\alpha) \leq d\}$ ordered by $\leq_{gl}$. $\mathbb{R}[X]$ 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 coefficients[1]. We use the same notation, $\deg(\cdot)$, to denote the degree of a polynomial or of an element of $\mathbb{N}^p$. For $d \in \mathbb{N}$, $\mathbb{R}_d[X]$ denotes the set of p-variate polynomials of degree less or equal to d. We set $s(d) = \binom{p+d}{d}$, the number of monomials of degree less or equal to d. We will denote by $\mathbf{v}_d(X)$ the vector of monomials of degree less or equal to d sorted by $\leq_{gl}$. We let $\mathbf{v}_d(X) := (X^\alpha)_{\alpha \in \mathbb{N}^p_d} \in \mathbb{R}_d[X]^{s(d)}$. With this notation, we can write a polynomial $P \in \mathbb{R}_d[X]$ as follows $P(X) = \langle \mathbf{p}, \mathbf{v}_d(X) \rangle$ for some real vector of coefficients $\mathbf{p} = (p_\alpha)_{\alpha \in \mathbb{N}^p_d} \in \mathbb{R}^{s(d)}$ ordered using $\leq_{gl}$. Given $\mathbf{x} = (x_i)_{i=1\dots p} \in \mathbb{R}^p$, $P(\mathbf{x})$ denotes the evaluation of P with the assignments $X_1 = x_1, X_2 = x_2, \dots X_p = x_p$. Given a Borel probability measure μ and $\alpha \in \mathbb{N}^p$, $y_\alpha(\mu)$ denotes the moment α of μ: $y_\alpha(\mu) = \int_{\mathbb{R}^p} \mathbf{x}^\alpha d\mu(\mathbf{x})$. Throughout the paper, we will only consider measures of which all moments are finite.

2.2 Moment matrix

Given a Borel probability measure μ on $\mathbb{R}^p$, the moment matrix of μ, $M_d(\mu)$, is a matrix indexed by monomials of degree at most d ordered by $\leq_{gl}$. For $\alpha, \beta \in \mathbb{N}^p_d$, the corresponding entry in $M_d(\mu)$ is defined by $M_d(\mu)_{\alpha,\beta} := y_{\alpha+\beta}(\mu)$, the moment $\alpha + \beta$ of μ. When $p = 2$, letting $y_\alpha = y_\alpha(\mu)$ for $\alpha \in \mathbb{N}^2_4$, we have

$$M_2(\mu): \begin{array}{r|cccccc} & 1 & X_1 & X_2 & X_1^2 & X_1X_2 & X_2^2 \\ \hline 1 & 1 & y_{10} & y_{01} & y_{20} & y_{11} & y_{02} \\ X_1 & y_{10} & y_{20} & y_{11} & y_{30} & y_{21} & y_{12} \\ X_2 & y_{01} & y_{11} & y_{02} & y_{21} & y_{12} & y_{03} \\ X_1^2 & y_{20} & y_{30} & y_{21} & y_{40} & y_{31} & y_{22} \\ X_1X_2 & y_{11} & y_{21} & y_{12} & y_{31} & y_{22} & y_{13} \\ X_2^2 & y_{02} & y_{12} & y_{03} & y_{22} & y_{13} & y_{04} \end{array}.$$

$M_d(\mu)$ is positive semidefinite for all $d \in \mathbb{N}$. Indeed, for any $\mathbf{p} \in \mathbb{R}^{s(d)}$, let $P \in \mathbb{R}_d[X]$ be the polynomial with vector of coefficients $\mathbf{p}$, we have $\mathbf{p}^T M_d(\mu) \mathbf{p} = \int_{\mathbb{R}^p} P^2(\mathbf{x}) d\mu(\mathbf{x}) \geq 0$. Furthermore, we have the identity $M_d(\mu) = \int_{\mathbb{R}^p} \mathbf{v}_d(\mathbf{x}) \mathbf{v}_d(\mathbf{x})^T d\mu(\mathbf{x})$ where the integral is understood elementwise.

2.3 Sum of squares (SOS)

We denote by $\Sigma[X] \subset \mathbb{R}[X]$ (resp. $\Sigma_d[X] \subset \mathbb{R}_d[X]$), the set of polynomials (resp. polynomials of degree at most d) which can be written as a sum of squares of polynomials. Let $P \in \mathbb{R}_{2m}[X]$ for some $m \in \mathbb{N}$, then P belongs to $\Sigma_{2m}[X]$ if there exists a finite $J \subset \mathbb{N}$ and a family of polynomials $P_j \in \mathbb{R}_m[X]$, $j \in J$, such that $P = \sum_{j \in J} P_j^2$. 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 $\Sigma_{2m}[X]$ if and only if

$$\exists Q \in \mathbb{R}^{s(m) \times s(m)},\ Q \succeq 0,\ P(\mathbf{x}) = \mathbf{v}_d(\mathbf{x})^T Q \mathbf{v}_d(\mathbf{x}),\ \forall \mathbf{x} \in \mathbb{R}^p. \tag{1}$$

As a consequence, every positive semidefinite matrix $Q \in \mathbb{R}^{s(m) \times s(m)}$ defines a polynomial in $\Sigma_{2m}[X]$ by using the representation in (1).

[1]For the null polynomial, we use the convention that its degree is 0 and it is $\leq_{gl}$ smaller than all other monomials.

3 Empirical observations on the inverse moment matrix SOS polynomial

The *inverse moment-matrix SOS polynomial* is associated to a measure μ which satisfies the following.

Assumption 1 *μ is a Borel probability measure on $\mathbb{R}^p$ with all its moments finite and $M_d(\mu)$ is positive definite for a given $d \in \mathbb{N}$.*

Definition 1 *Let μ, d satisfy Assumption 1. We call the SOS polynomial $Q_{\mu,d} \in \Sigma_{2d}[X]$ defined by the application:*

$$\mathbf{x} \mapsto \quad Q_{\mu,d}(\mathbf{x}) := \mathbf{v}_d(\mathbf{x})^T M_d(\mu)^{-1} \mathbf{v}_d(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^p, \tag{2}$$

the inverse moment-matrix SOS polynomial of degree $2d$ associated to μ.

Actually, connection to orthogonal polynomials will show that the inverse function $\mathbf{x} \mapsto Q_{\mu,d}(\mathbf{x})^{-1}$ is called the *Christoffel* function in the literature [13, 3] (see also Section 4).

In the remainder of this section, we focus on the situation when μ corresponds to an empirical measure over n points in $\mathbb{R}^p$ which are fixed. So let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be a fixed set of points and let $\mu := \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$ where $\delta_{\mathbf{x}}$ corresponds to the Dirac measure at $\mathbf{x}$. In such a case the polynomial $Q_{\mu,d}$ in (2) is determined only by the empirical moments up to degree $2d$ of our collection of points. Note that we also require that $M_d(\mu) \succ 0$. In other words, the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ 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 $\mathbb{R}^2$ for a fixed $d \in \mathbb{N}$: represent on the same graphic, the cloud of points $\{\mathbf{x}_i\}_{i=1\ldots n}$ and the sublevel sets of SOS polynomial $Q_{\mu,d}$ in $\mathbb{R}^2$ (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 $\{\mathbf{x} : Q_{\mu,d}(\mathbf{x}) \leq \binom{p+d}{d}\}$ captures most of the points. We could reproduce very similar observations on different shapes with various number of points in $\mathbb{R}^2$ and degree d (see Appendix A).

3.2 Measuring outlyingness

An additional remark in a similar line is that $Q_{\mu,d}$ 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 $Q_{\mu,d}$ here. Instead, the color and shape of the points are taken proportionally to the value of $Q_{\mu,d}$, 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 $Q_{\mu,d}$ 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 likelihood estimation for the Gaussian, up to a constant additive factor. To see this, set $\mu = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ and $S = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$. With this notation, we have

the following block representation of the moment matrix,

$$M_d(\mu) = \begin{pmatrix} 1 & \mu^T \\ \mu & S \end{pmatrix} \qquad M_d(\mu)^{-1} = \begin{pmatrix} 1 + \mu^T V^{-1} \mu & -\mu^T V^{-1} \\ -V^{-1}\mu & V^{-1} \end{pmatrix},$$

where $V = S - \mu\mu^T$ is the empirical covariance matrix and the expression for the inverse is given by Schur complement. In this case, we have $Q_{\mu,1}(\mathbf{x}) = 1 + (\mathbf{x} - \mu)^T V^{-1}(\mathbf{x} - \mu)$ for all $\mathbf{x} \in \mathbb{R}^p$. 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.

3.4 Computational aspects

Recall that $s(d) = \binom{p+d}{d}$ is the number of p-variate monomials of degree up to d. The computation of $Q_{\mu,d}$ 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 $Q_{\mu,d}$ requires $O(s(d)^2)$ operations.

Estimating the coefficients of $Q_{\mu,d}$ has a computational cost that depends only linearly in the number of points n. The cost of evaluating $Q_{\mu,d}$ 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 $O(n^2)$ operations for the evaluation of the model [1]. Moreover, this is well suited for online settings where inverse moment matrix computation can be done using Woodbury updates.

The dependence in the dimension p is of the order of p^d for a fixed d. Similarly, the dependence in d is of the order of d^p for a fixed dimension p and the joint dependence is exponential. This suggests that the computation and evaluation of $Q_{\mu,d}$ will mostly make sense for moderate dimensions and degree d.

4 Invariances 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 μ on $\mathbb{R}^p$ which satisfies Assumption 1. Note that $M_d(\mu)$ is always positive definite if μ is not supported on the zero set of a polynomial of degree at most d. Under Assumption 1, $M_d(\mu)$ induces an inner product on $\mathbb{R}^{s(d)}$ and by extension on $\mathbb{R}_d[X]$ (see Section 2). This inner product is denoted by $\langle \cdot, \cdot \rangle_\mu$ and satisfies for any polynomials $P, Q \in \mathbb{R}_d[X]$ with coefficients $\mathbf{p}, \mathbf{q} \in \mathbb{R}^{s(d)}$,

$$\langle P, Q \rangle_\mu := \langle \mathbf{p}, M_d(\mu)\mathbf{q} \rangle_{\mathbb{R}^{s(d)}} = \int_{\mathbb{R}^p} P(\mathbf{x})Q(\mathbf{x})d\mu(\mathbf{x}).$$

We will also use the canonical inner product over $\mathbb{R}_d[X]$ which we write $\langle P, Q \rangle_{\mathbb{R}_d[X]} := \langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{R}^{s(d)}}$ for any polynomials $P, Q \in \mathbb{R}_d[X]$ with coefficients $\mathbf{p}, \mathbf{q} \in \mathbb{R}^{s(d)}$. We will omit the subscripts for this canonical inner product and use $\langle \cdot, \cdot \rangle$ for both products.

4.1 Affine invariance

It is worth noticing that the mapping $\mathbf{x} \mapsto Q_{\mu,d}(\mathbf{x})$ does not depend on the particular choice of $\mathbf{v}_d(X)$ as a basis of $\mathbb{R}_d[X]$, any other basis would lead to the same mapping. This leads to the result that $Q_{\mu,d}$ captures affine invariant properties of μ.

Lemma 1 *Let μ satisfy Assumption 1 and $A \in \mathbb{R}^{p \times p}, b \in \mathbb{R}^p$ define an invertible affine mapping on $\mathbb{R}^p$, $\mathcal{A}\colon \mathbf{x} \to A\mathbf{x}+b$. Then, the push foward measure, defined by $\tilde{\mu}(S) = \mu(\mathcal{A}^{-1}(S))$ for all Borel sets $S \subset \mathbb{R}^p$, satisfies Assumption 1 (with the same d as μ) and for all $\mathbf{x} \in \mathbb{R}^p$, $Q_{\mu,d}(\mathbf{x}) = Q_{\tilde{\mu},d}(A\mathbf{x} + b)$.*

Lemma 1 is probably better understood when $\mu = 1/n \sum_{i=1}^n \delta_{\mathbf{x}_i}$ as in Section 3. In this case, we have $\tilde{\mu} = 1/n \sum_{i=1}^n \delta_{A\mathbf{x}_i + b}$ and Lemma 1 asserts that the level sets of $Q_{\tilde{\mu},d}$ are simply the images of those of $Q_{\mu,d}$ under the affine transformation $\mathbf{x} \mapsto A\mathbf{x} + b$. This is illustrated in Appendix D.

4.2 Connection with orthogonal polynomials

We define a classical [13, 3] family of orthonormal polynomials, $\{P_\alpha\}_{\alpha\in\mathbb{N}^p_d}$ ordered according to $\leq_{gl}$ which satisfies for all $\alpha \in \mathbb{N}^p_d$

$$\langle P_\alpha, X^\beta\rangle = 0 \text{ if } \alpha <_{gl} \beta,\ \langle P_\alpha, P_\alpha\rangle_\mu = 1,\ \langle P_\alpha, X^\beta\rangle_\mu = 0 \text{ if } \beta <_{gl} \alpha,\ \langle P_\alpha, X^\alpha\rangle_\mu > 0. \quad (3)$$

It follows from (3) that $\langle P_\alpha, P_\beta\rangle_\mu = 0$ if $\alpha \neq \beta$. Existence and uniqueness of such a family is guaranteed by the Gram-Schmidt orthonormalization process following the $\leq_{gl}$ order on the monomials, and by the positivity of the moment matrix, see for instance [3, Theorem 3.1.11].

Let $D_d(\mu)$ be the lower triangular matrix which rows are the coefficients of the polynomials P_α defined in (3) ordered by $\leq_{gl}$. It can be shown that $D_d(\mu) = L_d(\mu)^{-T}$, where $L_d(\mu)$ is the Cholesky factorization of $M_d(\mu)$. Furthermore, there is a direct relation with the inverse moment matrix as $M_d(\mu)^{-1} = D_d(\mu)^T D_d(\mu)$ [7, Proof of Theorem 3.1]. This has the following consequence.

Lemma 2 *Let μ satisfy Assumption 1, then $Q_{\mu,d} = \sum_{\alpha\in\mathbb{N}^p_d} P^2_\alpha$, where the family $\{P_\alpha\}_{\alpha\in\mathbb{N}^p_d}$ is defined by (3) and $\int_{\mathbb{R}^p} Q_{\mu,d}(\mathbf{x})d\mu(\mathbf{x}) = s(d)$.*

That is, $Q_{\mu,d}$ is a very specific and distinguished SOS polynomial, the sum of squares of the orthonormal basis elements $\{P_\alpha\}_{\alpha\in\mathbb{N}^p_d}$ of $\mathbb{R}_d(X)$ (w.r.t. μ). Furthermore, the average value of $Q_{\mu,d}$ with respect to μ 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 $\{P_\alpha\}_{\alpha\in\mathbb{N}^p_d}$ defined in (3) is the unique 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.

$$\begin{aligned}\min_{Q_\alpha,\theta_\alpha,\alpha\in\mathbb{N}^p_d}\ & \frac{1}{2}\int_{\mathbb{R}^p}\sum_{\alpha\in\mathbb{N}^p_d} Q_\alpha(\mathbf{x})^2 d\mu(\mathbf{x}) \\ \text{s.t. }\ & \mathbf{q}_{\alpha\alpha} \geq \exp(\theta_\alpha),\quad \mathbf{q}_{\alpha\beta} = 0,\quad \alpha,\beta\in\mathbb{N}^p_d,\quad \alpha <_{gl}\beta,\quad \sum_{\alpha\in\mathbb{N}^p_d}\theta_\alpha = 0,\end{aligned} \quad (4)$$

where $Q_\alpha(\mathbf{x}) = \sum_{\beta\in\mathbb{N}^p_d}\mathbf{q}_{\alpha\beta}\mathbf{x}^\beta$, $\alpha\in\mathbb{N}^p_d$. We first comment on problem (4). Let $P = \sum_{\alpha\in\mathbb{N}^p_d} Q^2_\alpha$ be the SOS polynomial appearing in the objective function of (4). The constraints of problem (4) restrict P to be in a certain set $S_d \subset \Sigma_d[X]$. With this notation, problem (4) is reformulated as $\min_{P\in S_d}\int P d\mu$. Therefore problem (4) balances two antagonist targets, on one hand the minimization of the average value of the SOS polynomial P with respect to μ, on the other hand the avoidance of the trivial polynomial, enforced by the constraint that $P \in S_d$. The constraints on P are simple and natural, they ensure that P is a sum of squares of polynomials $\{Q_\alpha\}_{\alpha\in\mathbb{N}^p_d}$, where the leading term of Q_α (according to the ordering $\leq_{gl}$) is $\mathbf{q}_{\alpha\alpha}\mathbf{x}^\alpha$ with $\mathbf{q}_{\alpha\alpha} > 0$ (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 $aQ \in S_d$. This suggests that S_d is a quite general class of nonvanishing SOS polynomials. The following result, which gives a relation between $Q_{\mu,d}$ and solutions of (4), uses a generalization of [13, 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 $(Q^*_\alpha, \theta^*_\alpha)$, which satisfies $Q^*_\alpha = \sqrt{\lambda}P_\alpha$, $\alpha\in\mathbb{N}^p_d$, for some $\lambda > 0$. In particular, the distinguished SOS polynomial $Q_{\mu,d} = \sum_{\alpha\in\mathbb{N}^p_d} P^2_\alpha = \frac{1}{\lambda}\sum_{\alpha\in\mathbb{N}^p_d}(Q^*_\alpha)^2$, is (part of) the unique optimal solution of (4).*

Theorem 1 states that up to the scaling factor λ, the distinguished SOS polynomial $Q_{\mu,d}$ 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 $\alpha \in \mathbb{N}^p_d$ (and again up to a scaling factor) the polynomial P_α is the unique optimal solution of the problem:

$\min_Q \left\{ \int Q^2 d\mu \,:\, Q \in \mathbb{R}_d[X],\, Q(\mathbf{x}) = \mathbf{x}^\alpha + \sum_{\beta <_{glex} \alpha} \mathbf{q}_\beta \, \mathbf{x}^\beta \right\}$. This fact is well-known in the univariate case [13, Theorem 3.1.2] and does not seem to have been exploited in the literature, at least for purposes similar to ours. So intuitively, P_α^2 should be as close to 0 as possible on the support of μ. Problem (4) has similar properties and the constraint on the vector of weights θ enforces that, at an optimal solution, the contribution $\int (Q_\alpha^*)^2 \, d\mu$ to the overall sum in the criterion is the same for all α. Using Lemma 2 yields (up to a multiplicative constant) the polynomial $Q_{\mu,d}$. Other constraints on θ would yield different weighted sum of the squares P_α^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, $Q_{\mu,d}$ should be as close to 0 as possible on average while remaining in a large class of nonvanishing SOS polynomials.

4.4 Christoffel function and outlier detection

The following result from [3, Theorem 3.5.6] draws a direct connection between $Q_{\mu,d}$ and the *Chritoffel* function (the right hand side of (5)).

Theorem 2 ([3]) *Let Assumption 1 hold and let $\bar{\mathbf{x}} \in \mathbb{R}^p$ be fixed, arbitrary. Then*

$$Q_{\mu,d}(\bar{\mathbf{x}})^{-1} \;=\; \min_{P \in \mathbb{R}_d[X]} \left\{ \int_{\mathbb{R}^p} P(\mathbf{x})^2 \, d\mu(\mathbf{x}) : P(\bar{\mathbf{x}}) \;=\; 1 \right\}. \tag{5}$$

Theorem 2 provides a mathematical rationale for the use of $Q_{\mu,d}$ for outlier or novelty detection purposes. Indeed, from Lemma 2 and equation (3), we have $Q_{\mu,d} \geq 1$ on $\mathbb{R}^p$. Furthermore, the solution of the minimization problem in (5) satisfies $P(\bar{\mathbf{x}})^2 = 1$ and $\mu\left(\left\{\mathbf{x} \in \mathbb{R}^p :\; P(\mathbf{x})^2 \leq 1\right\}\right) \geq 1 - Q_{\mu,d}(\bar{\mathbf{x}})^{-1}$ (by Markov's inequality). Hence, for high values of $Q_{\mu,d}(\bar{\mathbf{x}})$, the sublevel set $\left\{\mathbf{x} \in \mathbb{R}^p :\; P(\mathbf{x})^2 \leq 1\right\}$ contains most of the mass of μ while $P(\bar{\mathbf{x}})^2 = 1$. 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. For instance, the sub-level sets of $Q_{\mu,d}$, and in particular $\{\mathbf{x} \in \mathbb{R}^p : Q_{\mu,d}(\mathbf{x}) \leq \binom{p+d}{d}\}$, 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 $Q_{\mu,d}$ can be used to detect outliers in a collection of real vectors by taking μ to be the corresponding empirical measure. This is backed up by the results presented in Section 4. In this section we illustrate these properties on a real world example. We choose the KDD cup 99 network intrusion dataset (available at [11]) which consists of network connection data with labels describing whether they correspond to normal traffic or network intrusions. We follow [15] and [14] and construct five datasets consisting of labeled vectors in $\mathbb{R}^3$, the label indicating normal traffic or network attack. The content of these datasets is summarized in the following table.

Dataset	http	smtp	ftp-data	ftp	others
Number of examples	567498	95156	30464	4091	5858
Proportions of attacks	0.004	0.0003	0.023	0.077	0.016

The details on how these datasets are constructed are available in [15, 14] and are reproduced in Appendix C. The main idea is to give to each datapoint an outlyingness score solely based on its position in $\mathbb{R}^3$ and then compare outliers predicted by the score with the label indicating network intrusion. The underlying assumption is that network intrusion corresponds to infrequent abnormal behaviors and could thus be considered as outliers.

We reproduce the exact same experiment that was described in [14, Section 5.4] using the value of the inverse moment matrix SOS polynomial from Definition 1 as an outlyingness score (with $d = 3$). The authors of [14] have compared different types of methods for outlier detection in the same experimental setting: methods based on robust estimation and Mahalanobis distance [6, 8], mixture model based methods [12] and recurrent neural network based methods. These results are gathered

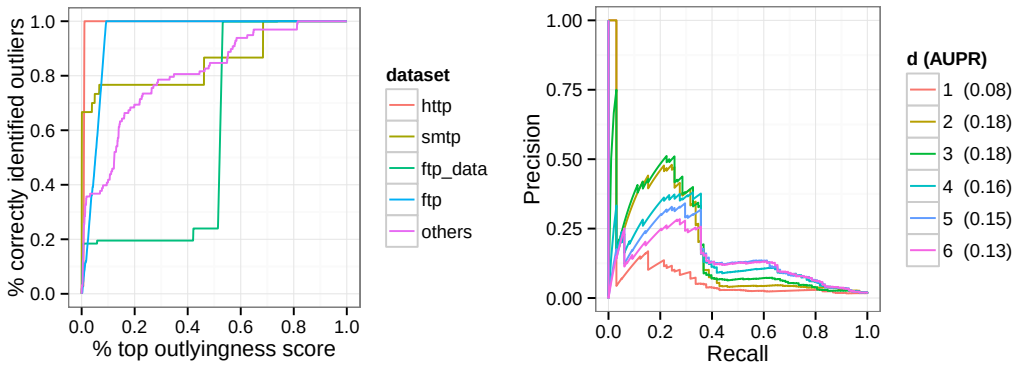

Figure 2: Left: reproduction of the results described in [14] with the inverse moment SOS polynomial value as an outlyingness score ($d = 3$). Right: Precision-recall curves for different values of d for the dataset "others".

in [14, Figure 7]. In the left panel of Figure 2 we represent the same performance measure for our approach. We first compute the value of the inverse moment SOS polynomial 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.

• 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 [14, 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 in $Q_{\mu,d}$ on the performances in terms of outlier detection. We focus on the "others" dataset because it is the most heterogeneous in term of data and outliers. We adopt a slightly different measure of performance and use precision recall curves (see for example [2]) 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 which controls the "complexity" of $Q_{\mu,d}$. Indeed, $2d$ is the degree of the polynomial $Q_{\mu,d}$ and it is expected that more complex models will potentially identify more diverse classes of examples of points as outliers. In our case, this means identifying regular traffic as outliers while it actually does not correspond to intrusions.

6 Conclusion and future work

We presented empirical findings with a mathematical intuition regarding the sublevel sets of the inverse moment matrix SOS polynomial. This opens many potential subjects of investigations.

- Similarities with maximum likelihood.
- Statistics in the context of empirical processes.
- Relation between a density and its inverse moment matrix SOS polynomial. Assymptotics when the degree increases.
- Connections with computational geometry and non Gaussian integrals.
- Computationally tractable extensions in higher dimensions.

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*.

References

[1] V. Chandola and A. Banerjee and V. Kumar (2009). Anomaly detection: A survey. ACM computing surveys (CSUR) 41(3):15.

[2] 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.

[3] C.F. Dunkl and Y. Xu (2001). Orthogonal polynomials of several variables. Cambridge University Press. MR1827871.

[4] 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).

[5] 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.

[6] 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.

[7] J.W. Helton and 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.

[8] E.M. Knorr and 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.

[9] J.B. Lasserre (2015). Level Sets and NonGaussian Integrals of Positively Homogeneous Functions. International Game Theory Review, 17(01):1540001.

[10] J.B. Lasserre and M.Putinar (2015). Algebraic-exponential Data Recovery from Moments. Discrete & Computational Geometry, 54(4):993-1012.

[11] M. Lichman (2013). UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml` University of California, Irvine, School of Information and Computer Sciences.

[12] J.J. Oliver and R.A.Baxter and C.S. Wallace (1996). Unsupervised learning using MML. Proceedings of the International Conference on Machine Learning (pp. 364-372).

[13] G. Szegö (1974). Orthogonal polynomials. In Colloquium publications, AMS, (23), fourth edition.

[14] G. Williams and R. Baxter and H. He and 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.

[15] K. Yamanishi and J.I. Takeuchi and 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.

A Additional examples

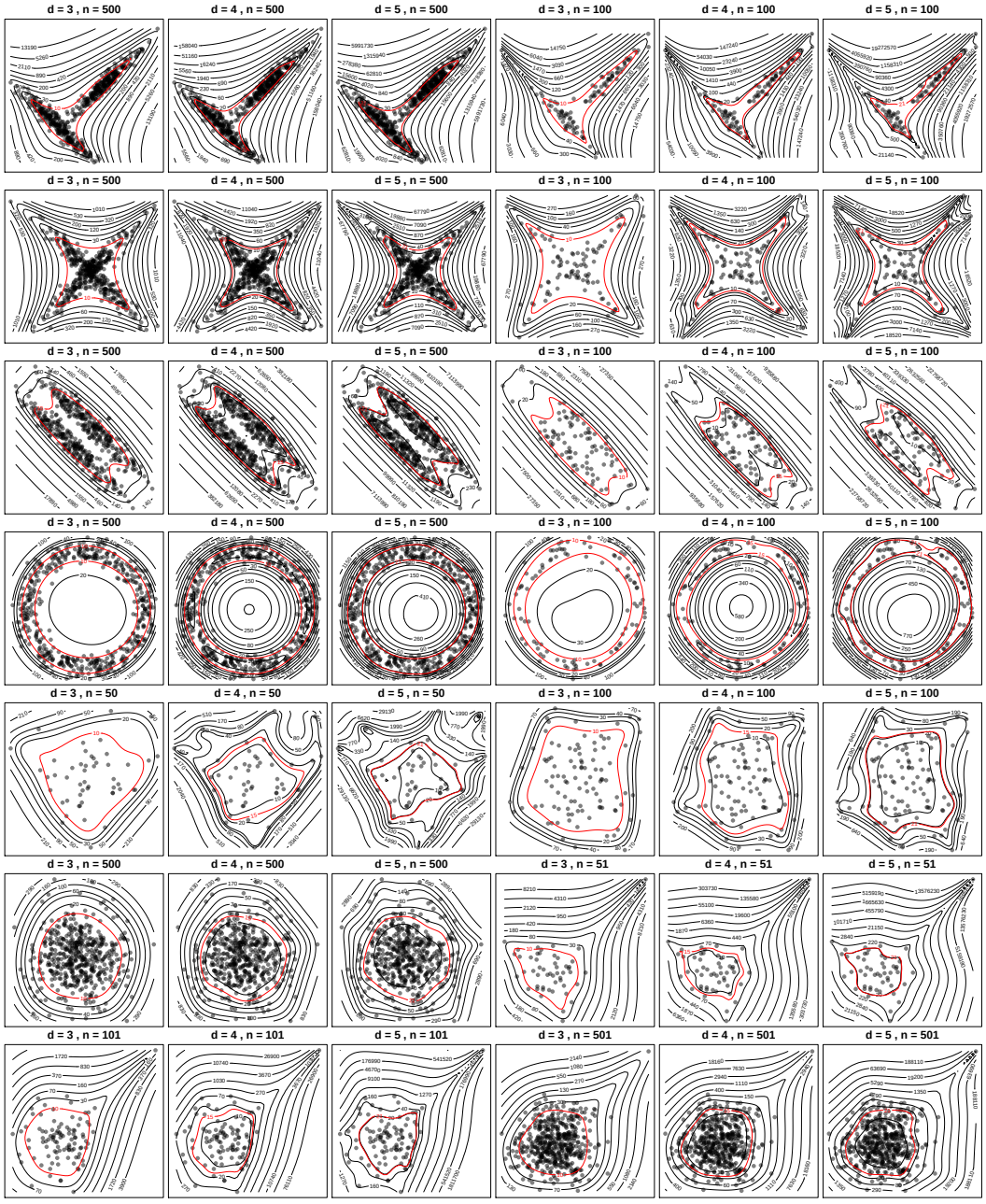

Figure 3: Empirical measure and level sets of $Q_{\mu,d}$ for various values of d and various configurations and numbers of points n. The level set $\binom{p+d}{d}$, which corresponds to the average value of $Q_{\mu,d}$, is represented in red.

B Proofs

We use the same notation as in the main text. We recall that Assumption 1.

Assumption 1 *μ is a Borel probability measure on $\mathbb{R}^p$ with all its moments finite and $M_d(\mu)$ is positive definite for a given $d \in \mathbb{N}$.*

Lemma 2 and Theorem 2 are taken from the literature and we provide a proof for completeness.

B.1 Proof of Lemma 1

First we show that the mapping $\mathbf{x} \mapsto Q_{\mu,d}(\mathbf{x})$ does not depend on the choice of a specific basis of $\mathbb{R}_d[X]$. Then we will deduce the affine invariance property.

Lemma 3 *Let $\mathbf{w}_d(X)$ be an arbitrary basis of $\mathbb{R}_d[X]$ and let $R_{\mu,d} \in \mathbb{R}_d[X]$ be derived in the same way as $Q_{\mu,d}$ (see Definition 1), with $\mathbf{w}_d$ in place of $\mathbf{v}_d$. Then $Q_{\mu,d}(\mathbf{x}) = R_{\mu,d}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^p$.*

Proof : Since $\mathbf{w}_d$ is a basis of $\mathbb{R}_d[X]$, there exists an invertible matrix $C \in \mathbb{R}^{s(d)\times s(d)}$ such that $\mathbf{w}_d(X) = C\mathbf{v}_d(X)$. We reproduce the computation of Definition 1 with this new basis. We write $N_d(\mu)$ the moment matrix computed with the polynomial basis $\mathbf{w}_d$. We have

$$\begin{aligned} N_d(\mu) &= \int_{\mathbb{R}^p} \mathbf{w}_d(\mathbf{x})\mathbf{w}_d(\mathbf{x})^T d\mu(\mathbf{x}) \\ &= \int_{\mathbb{R}^p} C\mathbf{v}_d(\mathbf{x})\mathbf{v}_d(\mathbf{x})^T C^T d\mu(\mathbf{x}) \\ &= C\int_{\mathbb{R}^p} \mathbf{v}_d(\mathbf{x})\mathbf{v}_d(\mathbf{x})^T d\mu(\mathbf{x}) C^T \\ &= CM_d(\mu)C^T, \end{aligned}$$

which leads to $N_d(\mu)^{-1} = C^{-T}M_d(\mu)^{-1}C^{-1}$. Using Definition 1, for all $\mathbf{x} \in \mathbb{R}^p$, we have

$$\begin{aligned} R_{\mu,d}(\mathbf{x}) &= \mathbf{w}_d(\mathbf{x})^T N_d(\mu)^{-1}\mathbf{w}_d(\mathbf{x}) \\ &= \mathbf{v}_d(\mathbf{x})^T C^T C^{-T} M_d(\mu)^{-1} C^{-1} C\mathbf{v}_d(\mathbf{x}) \\ &= \mathbf{v}_d(\mathbf{x})^T M_d(\mu)^{-1}\mathbf{v}_d(\mathbf{x}) \\ &= Q_{\mu,d}(\mathbf{x}), \end{aligned}$$

which concludes the proof. □

Lemma 1 *Let μ satisfy Assumption 1 and $A \in \mathbb{R}^{p\times p}, b \in \mathbb{R}^p$ define an invertible affine mapping on $\mathbb{R}^p$, $\mathcal{A}\colon \mathbf{x} \to A\mathbf{x}+b$. Then, the push foward measure, defined by $\tilde{\mu}(S) = \mu(\mathcal{A}^{-1}(S))$ for all Borel sets $S \subset \mathbb{R}^p$, satisfies Assumption 1 (with the same d as μ) and for all $\mathbf{x} \in \mathbb{R}^p$, $Q_{\mu,d}(\mathbf{x}) = Q_{\tilde{\mu},d}(A\mathbf{x}+b)$.*

Proof : Let us first compute $M_d(\tilde{\mu})$. For the push forward measure $\tilde{\mu}$, it holds that for any μ integrable function $f\colon \mathbb{R}^p \to \mathbb{R}$,

$$\int_{\mathbb{R}^p} f(\mathbf{x})d\tilde{\mu}(\mathbf{x}) = \int_{\mathbb{R}^p} f(A\mathbf{x}+b)d\mu(\mathbf{x}).$$

By considering polynomial f, we have that $\tilde{\mu}$ has all its moments finite and satisfies Assumption 1 with the same d as μ. Furthermore, we have

$$M_d(\tilde{\mu}) = \int_{\mathbb{R}^p} \mathbf{v}_d(\mathbf{x})\mathbf{v}_d(\mathbf{x})^T d\tilde{\mu}(\mathbf{x}) = \int_{\mathbb{R}^p} \mathbf{v}_d(A\mathbf{x}+b)\mathbf{v}_d(A\mathbf{x}+b)^T d\mu(\mathbf{x}). \tag{7}$$

We can deduce the following identity for all $\mathbf{x} \in \mathbb{R}^p$,

$$Q_{\tilde{\mu},d}(A\mathbf{x}+b) = \mathbf{v}_d(A\mathbf{x}+b)^T M_d(\tilde{\mu})^{-1}\,\mathbf{v}_d(A\mathbf{x}+b). \tag{8}$$

It remains to notice that mappings defined by $\mathbf{w}_d(\mathbf{x}) = \mathbf{v}_d(A\mathbf{x}+b)$ for all $\mathbf{x} \in \mathbb{R}^p$ form a basis of the polynomials of degree up to d on $\mathbb{R}^p$ (by invertibility of the affine mapping). Combining (7) and (8), we see that $\mathbf{x} \mapsto \mathbf{v}_d(A\mathbf{x}+b)$ simply corresponds to the use of a different basis of $\mathbb{R}_d[X]$. The result follows by applying Lemma 3 and the proof is complete. □

B.2 Proof of Lemma 2

Recall that the orthogonal polynomials satisfy for all $\alpha \in \mathbb{N}^p_d$

$$\langle P_\alpha, X^\beta \rangle = 0 \text{ if } \alpha <_{gl} \beta,\ \langle P_\alpha, P_\alpha \rangle_\mu = 1,\ \langle P_\alpha, X^\beta \rangle_\mu = 0 \text{ if } \beta <_{gl} \alpha,\ \langle P_\alpha, X^\alpha \rangle_\mu > 0. \tag{3}$$

Lemma 2 *Let μ satisfy Assumption 1, then $Q_{\mu,d} = \sum_{\alpha \in \mathbb{N}^p_d} P_\alpha^2$, where the family $\{P_\alpha\}_{\alpha \in \mathbb{N}^p_d}$ is defined by (3) and $\int_{\mathbb{R}^p} Q_{\mu,d}(\mathbf{x}) d\mu(\mathbf{x}) = s(d)$.*

Proof : Let $D_d(\mu)$ be the lower triangular matrix which rows are the coefficients of the polynomials P_α defined in (3) ordered by $\leq_{gl}$. From properties in (3), $D_d(\mu)$ is lower triangular with positive coefficients on its diagonal and therefore invertible. We have $D_d(\mu) M_d(\mu) D_d(\mu)^T = I$, the identity. It follows that $M_d(\mu) = D_d(\mu)^{-1} D_d(\mu)^{-T}$ and $M_d(\mu)^{-1} = D_d(\mu)^T D_d(\mu)$. Plugging this in definition 1 and using equation (1) leads to the desired identity. The average value result follows because we manipulate an orthonormal basis of $s(d)$ polymials, each of which has a square average value (with respect to μ) equal to 1. □

B.3 Proof of Theorem 1

We recall the the optimization problem.

$$\begin{aligned} \min_{Q_\alpha, \theta_\alpha, \alpha \in \mathbb{N}^p_d} \ & \frac{1}{2} \int_{\mathbb{R}^p} \sum_{\alpha \in \mathbb{N}^p_d} Q_\alpha(\mathbf{x})^2 d\mu(\mathbf{x}) \\ \text{s.t. } & \mathbf{q}_{\alpha\alpha} \geq \exp(\theta_\alpha), \quad \alpha \in \mathbb{N}^p_d, \\ & \mathbf{q}_{\alpha\beta} = 0,\ \alpha <_{gl} \beta, \quad \alpha, \beta \in \mathbb{N}^p_d, \\ & \sum_{\alpha \in \mathbb{N}^p_d} \theta_\alpha = 0. \end{aligned} \tag{4}$$

where $Q_\alpha(\mathbf{x}) = \sum_\beta \mathbf{q}_{\alpha\beta} \mathbf{x}^\beta$, $\alpha \in \mathbb{N}^p_d$. The statement of Theorem 1 goes as follows.

Theorem 1 *: Problem (4) is a convex optimization problem with a unique optimal solution $(Q^*_\alpha, \theta^*_\alpha)$, which satisfies $Q^*_\alpha = \sqrt{\lambda} P_\alpha$, $\alpha \in \mathbb{N}^p_d$, for some $\lambda > 0$. In particular, the distinguished SOS polynomial*

$$Q_{\mu,d} = \sum_{\alpha \in \mathbb{N}^p_d} P_\alpha^2 = \frac{1}{\lambda} \sum_{\alpha \in \mathbb{N}^p_d} (Q^*_\alpha)^2,$$

is (part of) the unique optimal solution of (4).

Proof :

General remarks. Observe that (4) is a convex optimization problem as we have $\int_{\mathbb{R}^p} \sum_{\alpha \in \mathbb{N}^p_d} Q_\alpha(\mathbf{x})^2 d\mu(\mathbf{x}) = \sum_{\alpha \in \mathbb{N}^p_d} \mathbf{q}_\alpha^T M_d(\mu)\, \mathbf{q}_\alpha$, which is strictly convex in $\{\mathbf{q}_\alpha\}_{\alpha \in \mathbb{N}^p_d}$. The proof is based on KKT optimality conditions for Problem (4). We first prove that any optimal solution should be of the form $Q^*_\alpha = \sqrt{\lambda} P_\alpha$, $\alpha \in \mathbb{N}^p_d$, for some $\lambda > 0$. Then we show that there exists a solution of the KKT system which has this form and finally that this solution is unique. The conclusion of Theorem 1 will then follow from Lemma 1. We begin with some notations that we will use throughout the proof.

Notation. Let $\{\mathbf{e}_\alpha\}_{\alpha \in \mathbb{N}^p_d}$ denote the canonical basis of $\mathbb{R}^{s(d)}$ indexed by $\alpha \in \mathbb{N}^p_d$ according to $\leq_{gl}$ order. The orthonormal polynomials $\{P_\alpha\}_{\alpha \in \mathbb{N}^p_d}$ (with respect to μ) are uniquely defined. For each $\alpha \in \mathbb{N}^p_d$, we write $\mathbf{p}_\alpha = (\mathbf{p}_{\alpha\beta})_{\beta \in \mathbb{N}^p_d} \in \mathbb{R}^{s(d)}$ the coefficients of the polynomial P_α. By construction of P_α, for every $\alpha, \beta \in \mathbb{N}^p_d$, $\alpha <_{gl} \beta$, $\mathbf{p}_{\alpha\beta} = 0$ and $\mathbf{p}_{\alpha\alpha} > 0$.

Optimality conditions Problem (4) is strictly feasible, we can choose any θ such that $\sum_\alpha \theta_\alpha = 0$ and for every $\alpha \in \mathbb{N}^p_d$, set $Q_\alpha := \kappa P_\alpha$ for some sufficiently large $\kappa > 0$. Therefore the KKT optimality conditions are necessary and sufficient for global optimality. We introduce Lagrange multipliers for problem (4): $\lambda_\alpha \geq 0$ for each inequality constraint, $\lambda_{\alpha\beta} \in \mathbb{R}$ for each linear equality constraint on polynomials with $\alpha <_{gl} \beta$ and $\lambda \in \mathbb{R}$ for the last linear equality constraint on $\{\theta_\alpha\}_{\alpha\in\mathbb{N}^p_d}$. The KKT optimality conditions for problem (4) can be written as follows

$$\lambda_\alpha \geq 0, \quad \mathbf{e}_\alpha^T \mathbf{q}^*_\alpha \geq \exp(\theta^*_\alpha), \quad \alpha \in \mathbb{N}^p_d, \tag{9}$$

$$\mathbf{e}_\beta^T \mathbf{q}^*_\alpha = 0, \quad \alpha, \beta \in \mathbb{N}^p_d, \quad \alpha <_{gl} \beta, \tag{10}$$

$$\sum_{\alpha\in\mathbb{N}^p_d} \theta^*_\alpha = 0, \tag{11}$$

$$M_d(\mu)\mathbf{q}^*_\alpha = \lambda_\alpha \mathbf{e}_\alpha + \sum_{\alpha <_{gl} \beta} \lambda_{\alpha\beta}\, \mathbf{e}_\beta, \quad \alpha \in \mathbb{N}^p_d, \tag{12}$$

$$\lambda_\alpha \exp(\theta^*_\alpha) = \lambda_\alpha \mathbf{e}_\alpha^T \mathbf{q}^*_\alpha = \lambda, \quad \alpha \in \mathbb{N}^p_d, \tag{13}$$

for optimal variables θ^*_α, polynomials Q^*_α with coefficients $\mathbf{q}^*_\alpha \in \mathbb{R}^{s(d)}$, for each $\alpha \in \mathbb{N}^p_d$. We next show that the part $(Q^*_\alpha)_{\alpha\in\mathbb{N}^p_d}$ of an optimal solution is necessarily a family of orthogonal polynomials.

Any optimal solution has the form $Q^*_\alpha = \sqrt{\lambda} P_\alpha$, $\alpha \in \mathbb{N}^p_d$, for some $\lambda > 0$. Since KKT conditions are necessary and sufficient for optimality, we only focus on them. For each $\alpha \neq 0$ and $\beta <_{gl} \alpha$, multiplying (12) by $\mathbf{e}_\beta$, we obtain

$$\left\langle X^\beta, Q^*_\alpha \right\rangle_\mu = \int \mathbf{x}^\beta Q^*_\alpha(\mathbf{x})\, d\mu(\mathbf{x}) \;=\; \mathbf{e}_\beta^T\, M_d(\mu)\, \mathbf{q}^*_\alpha \;=\; \lambda_\alpha\, \mathbf{e}_\beta^T \mathbf{e}_\alpha + \sum_{\alpha <_{gl} \gamma} \lambda_{\alpha\gamma}\, \mathbf{e}_\beta^T \mathbf{e}_\gamma \;=\; 0. \tag{14}$$

Similarly, multiplying (12) by $\mathbf{q}^*_\alpha$ yields for all $\alpha \in \mathbb{N}^p_d$,

$$\langle Q^*_\alpha, Q^*_\alpha \rangle_\mu = \int Q^*_\alpha(\mathbf{x})^2\, d\mu(\mathbf{x}) \;=\; (\mathbf{q}^*_\alpha)^T\, M_d(\mu)\, \mathbf{q}^*_\alpha \;=\; \lambda_\alpha\, (\mathbf{q}^*_\alpha)^T \mathbf{e}_\alpha \;=\; \lambda, \tag{15}$$

where we have used (13) for the last identity. In particular, with $\alpha = 0$, $Q^*_0(\mathbf{x}) = \mathbf{q}^*_{00}$ ($\geq \exp(\theta^*_0)$) for all $\mathbf{x}$ and so

$$\lambda \;=\; \int Q^*_0(\mathbf{x})^2\, d\mu(\mathbf{x}) \;=\; (\mathbf{q}^*_{00})^2 \int d\mu \geq \exp(2\theta^*_0),$$

which shows that $\lambda > 0$. Next, combining (14), (15) and the condition (10), we immediately deduce

$$\left\langle Q^*_\beta, Q^*_\alpha \right\rangle_\mu = \int Q^*_\beta(\mathbf{x})\, Q^*_\alpha(\mathbf{x})\, d\mu(\mathbf{x}) = \begin{cases} \lambda & \text{if } \alpha = \beta \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Finally, for every $\alpha \in \mathbb{N}^p_d$, multiplying (12) by $\mathbf{e}_\alpha$ yields

$$\langle X^\alpha, Q^*_\alpha \rangle_\mu = \int \mathbf{x}^\alpha Q^*_\alpha(\mathbf{x})\, d\mu(\mathbf{x}) \;=\; \mathbf{e}_\alpha^T\, M_d(\mu)\, \mathbf{q}^*_\alpha \;=\; \lambda_\alpha > 0, \quad \alpha \in \mathbb{N}^p_d. \tag{17}$$

The last inequality follows from (13). Indeed, suppose $\lambda_\alpha = 0$ for some $\alpha \in \mathbb{N}^p_d$, this would yield $\lambda = 0$. Since we have shown that $\lambda > 0$, it must also hold that $\lambda_\alpha > 0$ for all α. Combining relations (10), (14), (16) and (17), we have shown that the $\{Q^*_\alpha\}_{\alpha\in\mathbb{N}^p_d}$ form a family of orthogonal polynomials with respect to μ. In addition, by the uniqueness of the orthonormal basis $\{P_\alpha\}_{\alpha\in\mathbb{N}^p_d}$, it follows from (16) that $Q^*_\alpha = \sqrt{\lambda}\, P_\alpha$ for every $\alpha \in \mathbb{N}^p_d$.

There exists a solution of this form. Recall that, for each $\alpha \in \mathbb{N}^p_d$, $\mathbf{p}_\alpha = (\mathbf{p}_{\alpha\beta})_{\beta\in\mathbb{N}^p_d} \in \mathbb{R}^{s(d)}$ is the vector of coefficients of the polynomial P_α which satisfies by construction $\mathbf{p}_{\alpha\alpha} > 0$ and $\mathbf{p}_{\alpha\beta} = 0$

for all $\beta \in \mathbb{N}^p_d$, $\alpha <_{gl} \beta$. We use the following assignment for the primal and dual variables.

$$\lambda = \left(\prod_{\alpha \in \mathbb{N}^p_d} \mathbf{p}_{\alpha\alpha} \right)^{\frac{-2}{s(d)}} > 0 \tag{18}$$

$$\begin{aligned}
\lambda_\alpha &= \sqrt{\lambda} \mathbf{e}_\alpha^T M_d(\mu) \mathbf{p}_\alpha = \frac{\sqrt{\lambda}}{\mathbf{p}_{\alpha\alpha}} > 0, \quad \alpha \in \mathbb{N}^p_d \\
\lambda_{\alpha\beta} &= \sqrt{\lambda} \mathbf{e}_\beta^T M_d(\mu) \mathbf{p}_\alpha, \quad \alpha, \beta \in \mathbb{N}^p_d, \alpha <_{gl} \beta \\
\mathbf{q}^*_\alpha &= \sqrt{\lambda} \mathbf{p}_\alpha, \quad \theta^*_\alpha = \log(\sqrt{\lambda} \mathbf{p}_{\alpha\alpha}), \quad \alpha \in \mathbb{N}^p_d.
\end{aligned}$$

Using orthonormality of the polynomials $\{P_\alpha\}_{\alpha \in \mathbb{N}^p_d}$, it can be check that the assignment (18) satisfies KKT optimality conditions (9), (10), (11), (12) and (13). We have therefore constructed an optimal solution of (4) with the desired form.

The optimal solution is unique. From what precedes any optimal solution of (4) is necessarily such that $Q^*_\alpha = \sqrt{\lambda} P_\alpha$, for every $\alpha \in \mathbb{N}^n$, for some $\lambda > 0$. In addition the optimal value of (4) is $s(d)\lambda$. Suppose that there exists two different optimal solutions $(Q_\alpha, \theta_\alpha)_{\alpha \in \mathbb{N}^p_d}$ and $(Q'_\alpha, \theta'_\alpha)_{\alpha \in \mathbb{N}^p_d}$ with associated dual variables $(\lambda, \lambda_\alpha, \lambda_{\alpha\beta})_{\alpha,\beta \in \mathbb{N}^p_d}$ and $(\lambda', \lambda'_\alpha, \lambda'_{\alpha\beta})_{\alpha,\beta \in \mathbb{N}^p_d}$. Then necessarily $\lambda = \lambda'$, $Q_\alpha = Q'_\alpha = \sqrt{\lambda} P_\alpha$ and $\lambda_\alpha, \lambda'_\alpha > 0$ for all $\alpha \in \mathbb{N}^p_d$. But then from (13), $\sqrt{\lambda} \mathbf{p}_{\alpha\alpha} = \exp(\theta_\alpha) = \exp(\theta'_\alpha)$ and so $\theta'_\alpha = \theta_\alpha$ for every $\alpha \in \mathbb{N}^p_d$. Therefore the solution is unique and this concludes the proof of Theorem 1. □

B.4 Proof of Theorem 2

Theorem 2 *Let Assumption 1 hold and let $\bar{\mathbf{x}} \in \mathbb{R}^p$ be fixed, arbitrary. Then*

$$Q_{\mu,d}(\bar{\mathbf{x}})^{-1} \;=\; \min_{P \in \mathbb{R}_d[X]} \left\{ \int P(\mathbf{x})^2 \, d\mu : P(\bar{\mathbf{x}}) \;=\; 1 \right\}. \tag{7}$$

Proof :

Fix an arbitrary $P \in \mathbb{R}_d[X]$ and $\bar{\mathbf{x}} \in \mathbb{R}^p$. Assume that $P(\bar{\mathbf{x}}) = 1$. Letting for all $\alpha \in \mathbb{N}^p_d$, $a_\alpha = \langle P, P_\alpha \rangle_\mu$, by orthonormality, we have

$$\begin{aligned}
P &= \sum_{\alpha \in \mathbb{N}^p_d} a_\alpha P_\alpha, \\
\langle P, P \rangle_\mu &= \sum_{\alpha \in \mathbb{N}^p_d} a^2_\alpha.
\end{aligned} \tag{19}$$

The assumption that $P(\bar{\mathbf{x}}) = 1$ can be used in conjonction with Cauchy-Schwartz inequality to obtain

$$\begin{aligned}
1 &= P(\bar{\mathbf{x}}) \\
&= \sum_{\alpha \in \mathbb{N}^p_d} a_\alpha P_\alpha(\bar{\mathbf{x}}) \\
&\leq \left(\sum_{\alpha \in \mathbb{N}^p_d} a^2_\alpha \right) \left(\sum_{\alpha \in \mathbb{N}^p_d} P_\alpha(\bar{\mathbf{x}})^2 \right) \\
&= \langle P, P \rangle_\mu \, Q_{\mu,d}(\bar{\mathbf{x}}),
\end{aligned} \tag{20}$$

where the last equality comes from the definition of $\langle \cdot, \cdot \rangle_\mu$ and Lemma 2. There is equality in equation 20 if and only if $a_\alpha = P_\alpha(\bar{\mathbf{x}}) / Q_{\mu,d}(\bar{\mathbf{x}})$ which always leads to $P(\bar{\mathbf{x}}) = 1$. This shows that the infimum is attained and concludes the proof. □

C Details about the preparation of the datasets

We reproduce the exact same manipulations as in the references. We downloaded the `kddcup.data` from the following repository

`https://archive.ics.uci.edu/ml/machine-learning-databases/kddcup99-mld/`

This file contains 4898431 instances of network connections described by 42 features including the type of connection (attack or normal). We filter the records by keeping only those for which the variable *logged in* is positive. We kept the labels (type of connection) together with the four most important features: *service, duration, src_bytes, dst_bytes*. We applied to the three last variables (numerical) the function $\log(\cdot + 0.1)/10$. We build four datasets with the four most frequent instances of *service* and group all the remaining records in the dataset *others* to get our five datasets.

D Illustration of affine invariance

The following Figure illustrate the affine invariance property described in Lemma 1.

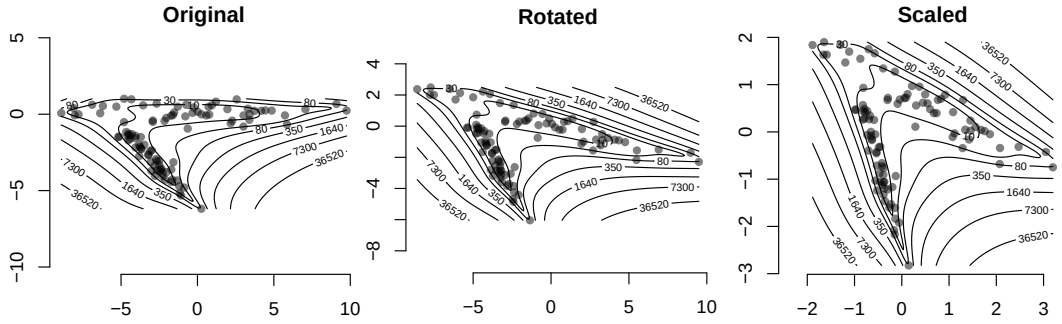

Figure 4: Empirical measure and level sets of $Q_{\mu,d}$ for three configurations of the same cloud of points ($d = 4$). The cloud in the middle is rotation of the original one and the cloud on the right is the same after centering and scaling. We observe that the level sets follow the same transformations.