2019 American Control Conference (ACC)
Philadelphia, PA, USA, July 10-12, 2019

A Sum of Squares Optimization Approach to Uncertainty Quantification

Brendon K. Colbert[1], Luis G. Crespo[2], and Matthew M. Peet[1].

Abstract— **This paper proposes a Sum of Squares (SOS) optimization technique for using multivariate data to estimate the probability density function of a non-Gaussian generating process. The class of distributions over which we optimize, result from using a polynomial map to lift the data into a higher-dimensional space, solving for an optimal Gaussian fit in this space, and then projecting a polynomial slice of the resulting joint density into physical space. The resulting distribution, to be called *Sliced Normal*, is able to characterize multimodal responses and strong parameter dependencies. We investigate several formulations of the problem, first maximizing a log-likelihood function, then a worst-case log-likelihood function, and finally using a heuristic to increase sparsity within the maximum log-likelihood formulation - thereby identifying independent subsets of the random variables. Using the optimal density functions in each scenario, we then propose semi-algebraic sets representing confidence regions, or "safe sets," for future data. Finally, we show numerically that these "safe sets" are reliable and, hence, can be used for system identification, fault detection, robustness analysis, and robust control design.**

I. Introduction

The characterization of the uncertainty in measured data - often representing model parameters - is of significance for system identification, robust analysis, and robust controller synthesis. Such variability arises from aleatory variation in physical parameters, varying operating conditions, model-form uncertainty, and measurement error. Here we tackle the problem of estimating the distribution of an unknown multivariate data-generating process based on samples, and calculating semi-algebraic representations of minimal "safe" regions of the parameter space where future data is likely to fall. The accurate characterization of such parameter regions can then be used to reduce the conservatism associated with the more common use of ellipsoidal representation of uncertainty [1]. Examples of work which use semialgebraic representations of parametric uncertainty for control include stability analysis of a model of Bacterial Heat Shock Response [2], and methods for calculating upper bounds on $\mathcal{H}_\infty$ and $\mathcal{H}_2$ system norms of linear systems [3].

Consider a data-generating process for the continuous parameter $\delta \in \mathbb{R}^n$ having an unknown structure. The main goal of this article is to characterize the underlying distribution of this process given the data sequence $\mathcal{D} = \{\delta^{(1)}, \dots, \delta^{(m)}\}$ comprised of IID samples.

For any $\mu \in \mathbb{R}^n$ and positive matrix $P \succeq 0 \in \mathbb{R}^{n\times n}$, the joint Probability Density Function (PDF) of a Gaussian is

$$f_G(\delta; \mu, P) = \frac{e^{-\frac{(\delta-\mu)^T P(\delta-\mu)}{2}}}{(2\pi)^{n/2}\sqrt{|P^{-1}|}}. \tag{1}$$

However, confidence regions associated with such a Gaussian will necessarily be ellipsoidal. Furthermore, Gaussian distributions fail to accurately describe skewed and multimodal distributions that often occur in practice, e.g., see the data in Figure 1 or 2. For this reason, we propose to parameterize a set of Gaussian PDFs: not in the original parameter space $\delta \in \mathbb{R}^n$, but in a lifted space $\mathbb{R}^q$ where $q > n$ and the map from $\mathbb{R}^n \to \mathbb{R}^q$ is given by $Z_d(\delta)$, where Z_d is the length-q vector of monomials in variables δ of degree greater than 0 and less than d, where $q = \binom{n+d}{n} - 1$. We may then define our class of generalized non-Gaussian PDFs as follows, for any $\mu \in \mathbb{R}^q$ and positive matrix $P \succeq 0 \in \mathbb{R}^{q\times q}$, we define a "candidate" PDF of the form

$$f(\delta; \mu, P) = \frac{e^{-\frac{(Z_d(\delta)-\mu)^\top P(Z_d(\delta)-\mu)}{2}}}{(2\pi)^{q/2}\sqrt{|P^{-1}|}}. \tag{2}$$

Naturally, when $d = 1$, $f(\delta; \mu, P)$ reduces to the Gaussian Normal distribution in $\mathbb{R}^n$. For $d > 1$, however, f is not a Gaussian distribution. Moreover, for a given PDF of the form $f(\delta; \mu, P)$, while the generated random variables δ do not have a Gaussian distribution, if we define a new set of random variables, defined as $z = Z_d(\delta)$, then these new variables *will* have a Gaussian distribution. Note that "candidate" PDFs of the form $f(\delta; \mu, P)$ do not integrate to 1 and hence require normalization in the original parameter space - a topic which will be addressed. We refer to these candidate distributions as "Sliced Normals".

Now that we have defined our objective, we briefly introduce our approach. Specifically, we need to define an optimization problem which determines the function $f(\delta; \mu, P)$ that best fits the data. This problem is complicated by the fact that $f(\delta; \mu, P)$ is a nonlinear function of P. For this reason we turn to the class of optimization problem defined by a log-likelihood objective function. Our first approach is to maximize the log-likelihood in the lifted space $\mathbb{R}^q$, of the given finite collection of data points $\mathcal{D} \subset \mathbb{R}^n$. In this case the objective function is given as the log of

$$L_f(\mathcal{D}) = \prod_{\delta \in \mathcal{D}} f(\delta; \mu, P). \tag{3}$$

Maximum likelihood approaches have been covered extensively in work such as [4]. As applied to our formulation, this approach has the advantage that the objective function becomes convex in the variables P and furthermore, the computational complexity of the resulting optimization problem

This work was supported by the NASA NIFS program and NSF grants CNS-1739990, and CMMI-1538374.

[1] Arizona State University, Tempe, AZ, 85287, USA. Brendon Colbert, bkcolbe1@asu.edu, is the corresponding author.

[2] NASA Langley Research Center, Hampton, VA, 23681, USA

978-1-5386-7928-9/$31.00 ©2019 AACC

does not depend on the size of the data set. Indeed, it can be shown that there is an analytic solution to this problem.

Unfortunately, we found that while the use of a log-likelihood objective yielded accurate representations of the 66% confidence regions of the underlying process, the more restrictive 99% confidence regions were too conservative - See Figure 2. For this reason, we propose an alternative formulation of the problem based on worst-case likelihood. Specifically, we propose to maximize the minimum likelihood of any point, $\delta \in \mathcal{D}$, evaluated in the lifted space $\mathbb{R}^q$. Thus, the second method maximizes the log of the function

$$W_f(\mathcal{D}) = \min_{\delta\in\mathcal{D}} f(\delta;\mu,P). \tag{4}$$

Empirically, we found that this objective function produced confidence regions of smaller volume, leading to less conservatism in the representation of safe parameter regions.

In both cases, we use the resulting optimal PDF to obtain semialgebraic representations of nested confidence regions of the form

$$S(\beta) = \{\delta : (Z_d(\delta)-\mu)^\top P(Z_d(\delta)-\mu) \le \beta\}, \tag{5}$$

where, for a desired percentage α, β must be numerically calculated (in a manner to be defined) such that α percent of the data is contained in $S(\beta)$. To analyze performance of the proposed algorithms, we then use newly generated test sets to determine the percentage of new data points generated by $G(\delta)$ which are contained in a given confidence region. These results are shown in Table I along with a metric for volume of the resulting sets.

Finally, in Section V, we propose a heuristic for identification of independent variables within the log-likelihood optimization framework. Identifying independent variables allows us to, e.g. decouple independent parameters - thereby reducing the complexity of the associated robust control problem. Our approach to identification of independent subsets of the data is to note that in our formulation P is the inverse of the covariance matrix Σ of the Gaussian PDF in the lifted q-dimensional space. Therefore, if we add a weight to the objective function which rewards block-diagonal structure of P, this will likewise result in a block-diagonal structure to the covariance matrix Σ. For this reason, we use an L_1 constraint on the off-diagonal terms of P and show that this increases the sparsity of P, thereby eliminating weak parameter dependencies in the resulting distribution- See Figure 4.

II. Notation

Denote by $\mathbb{S}^n$ and $\mathbb{S}^{n+}$ the symmetric matrices and cone of positive semi-definite matrices of size $n \times n$ respectively. Furthermore, let the function $Z_d : \mathbb{R}^n \to \mathbb{R}^q$ denote the vector of monomials of degree less than d but greater than 0, where $q = \binom{n+d}{n} - 1$. Finally, we denote the ring of multivariate polynomials with real coefficients as $\mathbb{R}[\delta]$.

III. Maximum Log Likelihood

In this section, we formulate the max log-likelihood optimization problem and show it can be reformulated as a Semidefinite Programming Problem (SDP) with an objective function which includes $\max \log|P|$. Such optimization problems can then be solved using SDP solvers such as SDPT3. Specifically, the log-likelihood optimization problem is formulated as

$$\max_{P\in\mathbb{S}^+,\ \mu\in\mathbb{R}^q}\left\{\log\prod_{\delta\in\mathcal{D}}\frac{e^{-\frac{(Z_d(\delta)-\mu)^\top P(Z_d(\delta)-\mu)}{2}}}{(2\pi)^{q/2}\sqrt{|P^{-1}|}} : P\succeq 0\right\}. \tag{6}$$

This optimization problem is a special case of optimization problems of the form

$$\max_{P\in\mathbb{S}^+,\ \mu\in\mathbb{R}^q}\left\{\log\prod_{i=1}^{m}\frac{e^{-\frac{(h_i-\mu)^\top P(h_i-\mu)}{2}}}{c\sqrt{|P^{-1}|}} : P\succeq 0\right\}. \tag{7}$$

Such optimization problems admit an analytic solution, as can be found in, e.g. [5]. Specifically, for a given data sequence $\{h^{(i)}\}_{i=1}^m$, the optimum is $\mu^* = \frac{1}{m}\sum_{i=1}^m h^{(i)}$ and $P^* = \Sigma^{-1}$, where $\Sigma = \frac{1}{m}\sum_{i=1}^m (h^{(i)}-\mu^*)(h^{(i)}-\mu^*)^\top$. However, as we will see, maximizing log likelihood by itself, even in a lifted space, does not result in an ideal fit to the data. Furthermore, this analytic solution cannot be readily modified for new objectives, additional constraints, or regularity conditions. For this reason, we treat the log likelihood problem explicitly in the optimization framework and do not rely on the existence of analytic solutions.

To define the optimization more precisely, we use an indexed data set $\mathcal{D} = \{\delta^{(i)}\}_{i=1}^m$, and define the lifted data points as $z_i = Z_d(\delta^{(i)}) - \mu$ where we *will* use $\mu = \mu^*$ as indicated in the analytic solution. Then we have that

$$\begin{aligned}
&\log\prod_{\delta\in\mathcal{D}}\frac{e^{-\frac{(Z_d(\delta)-\mu)^\top P(Z_d(\delta)-\mu)}{2}}}{(2\pi)^{q/2}\sqrt{|P^{-1}|}},\\
&= \log\left(\prod_{i=1}^m \frac{1}{(2\pi)^{q/2}\sqrt{|P^{-1}|}}\mathrm{e}^{-\frac{1}{2}z_i^\top P z_i}\right),\\
&= \sum_{i=1}^m \log\left(\frac{1}{(2\pi)^{q/2}\sqrt{|P^{-1}|}}\mathrm{e}^{-\frac{1}{2}z_i^\top P z_i}\right),\\
&= m\log\frac{1}{(2\pi)^{q/2}\sqrt{|P^{-1}|}} - \sum_{i=1}^m \frac{1}{2}z_i^\top P z_i,\\
&= -m\log((2\pi)^{q/2}) + \frac{m}{2}\log|P| - \frac{1}{2}\sum_{i=1}^m z_i^\top P z_i.
\end{aligned}$$

Since the leading term is independent of the decision variable P, the optimal P^* for Optimization Problem (6) is given by

$$P^* = \arg\max_{P\in\mathbb{S}^+}\left\{m\log|P| - \sum_{i=1}^m z_i^\top P z_i :\ P\succeq 0\right\}. \tag{8}$$

This problem may be further simplified if we define $z_i(j)$ be the j'th element of the vector $z_i \in \mathbb{R}^q$. Then the optimization

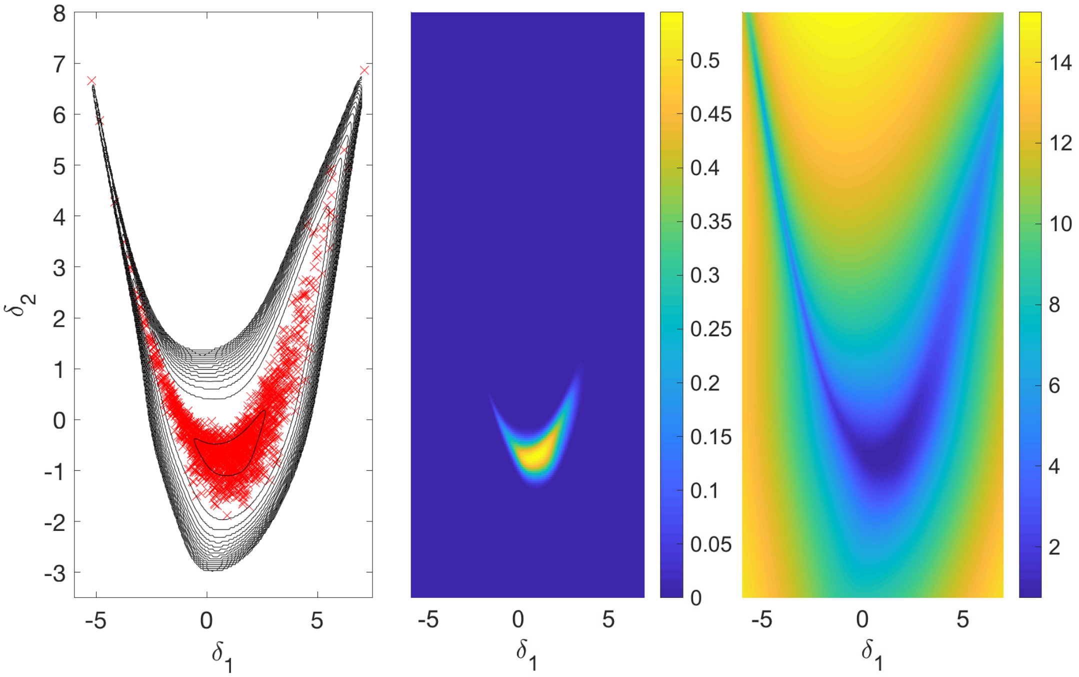

Fig. 1. Fig. a (left) - Data (red crosses) and level sets of $f(\delta;\mu,P)$, Fig. b (middle) - upper view of the PDF of the Sliced Normal $f(\delta;\mu,P)$, and Fig. c (right) - value of $Z_d(\delta)^\top P^* Z_d(\delta)$.

problem becomes

$$P^* = \arg\max_{P\in\mathbb{S}^+}\left\{ m\log|P| - \sum_{j,k=1}^{q}\sum_{i=1}^{m} z_i(j)z_i(k)P_{j,k} : P \succeq 0\right\}, \tag{9}$$

which is the combination of a $\log\det P$ term and a linear combination of the elements of P. Problems consisting of the log determinant of a positive semi-definite matrix P, and a linear combination of elements in P are convex with respect to the decision variable P [6] and may be solved efficiently with a suitable semi-definite optimization solver such as SDPT3 [7].

Example 1: We consider a data set $\mathcal{D} = \{\delta^{(i)}\}_{i=1}^m$, where $m = 500$, drawn from an unknown data-generating process. We solved Optimization Problem (8) using $d = 5$. Figure 1a shows the distribution of the data, along with level sets of the function $f(\delta;\mu,P)$. Clearly this data would be poorly represented using a Gaussian fit with associated ellipsoidal confidence regions. The proposed algorithm, meanwhile, is able to accurately capture the strong dependency between δ_1 and δ_2. Figures 1b and 1c show the upper view of the Sliced Normal PDF and of the argument of the exponential.

A. Computational Complexity Analysis

Optimization Problem (8) has an objective function whose number of constraints and number of variables are independent of the number of data points. This means that the optimization problem is dependent solely on the degree of the monomial basis, d, and the number of parameters, n.

In Fig. 3 we see the computation time for Optimization Problem (8) for 1000 data points and several different monomial degrees. Even for a degree 4 monomial basis we see that the optimization problem can be completed in well under one second.

IV. Worst-case Log-Likelihood

In this section, we formulate the worst-case log-likelihood optimization problem and show it can be reformulated as an SDP with an objective function that includes $\max\log|P|$. Such optimization problems can then be solved using standard SDP solvers such as SDPT3.

The worst-case log-likelihood optimization problem is then formulated as

$$\max_{P\in\mathbb{S}^+,\,\mu\in\mathbb{R}^q}\left\{ \min_{\delta\in\mathcal{D}} \log\frac{e^{-\frac{(Z_d(\delta)-\mu)^\top P(Z_d(\delta)-\mu)}{2}}}{(2\pi)^{q/2}\sqrt{|P^{-1}|}} : P \succeq 0\right\}. \tag{10}$$

As in the log-likelihood case, this problem can be simplified. Specifically, for a given indexed data set $\mathcal{D} = \{\delta^{(i)}\}_{i=1}^m$, define the lifted data points as $z_i = Z_d(\delta^{(i)}) - \mu$, where we must fix our variables μ. In this case, there is no analytic solution. Therefore, in order to convexify the problem, we relax the structure of the density slightly:

$$f(z;P) := \frac{e^{-\frac{\begin{bmatrix}1\\z\end{bmatrix}^\top P\begin{bmatrix}1\\z\end{bmatrix}}{2}}}{(2\pi)^{q/2}\sqrt{|P^{-1}|}}.$$

Then P is now of dimension $(q+1)\times(q+1)$ and we have

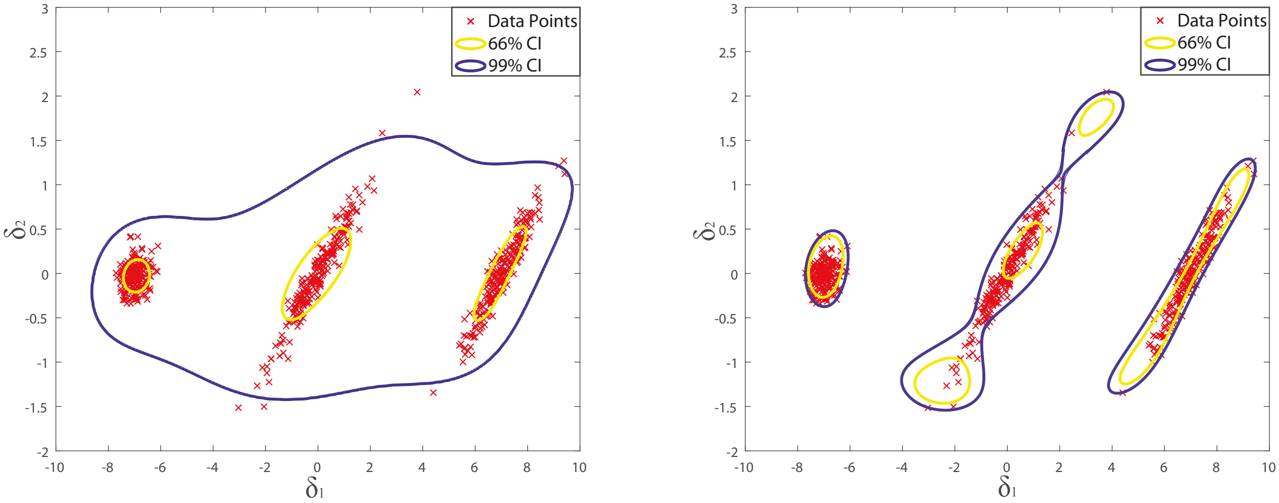

Fig. 2. Data from three separate bivariate Gaussian distributions whose mean in δ_1 is zero and in δ_2 is $-7, 0$ and 7 respectively. Semi-algebraic sets enclosing 66% and 99% of the generated data are superimposed, where Fig. 2(a) corresponds to the maximum likelihood formulation while Fig. 2(b) corresponds to the worst-case likelihood formulation.

that

$$\min_{\delta\in\mathcal{D}} \log\left(\frac{1}{(2\pi)^{q/2}\sqrt{|P^{-1}|}}\mathrm{e}^{-\frac{1}{2}\left[\begin{smallmatrix}1\\ z_d(\delta)\end{smallmatrix}\right]^\top P\left[\begin{smallmatrix}1\\ z_d(\delta)\end{smallmatrix}\right]}\right), \quad (11)$$

$$= \min_i\left(\log\frac{1}{(2\pi)^{q/2}\sqrt{|P^{-1}|}} - \frac{1}{2}\left[\begin{smallmatrix}1\\ z_i\end{smallmatrix}\right]^\top P\left[\begin{smallmatrix}1\\ z_i\end{smallmatrix}\right]\right),$$

$$= \min_i\left(-\log((2\pi)^{q/2}) + \frac{1}{2}\log|P| - \frac{1}{2}\left[\begin{smallmatrix}1\\ z_i\end{smallmatrix}\right]^\top P\left[\begin{smallmatrix}1\\ z_i\end{smallmatrix}\right]\right).$$

As in the log likelihood case, since the leading term is independent of the decision variable P, we may find P^* which solves Optimization Problem (11) by solving the following simplified optimization problem.

$$P^* = \arg\max_{P\in\mathbb{S}^+}\left\{t : t \le \log|P| - \left[\begin{smallmatrix}1\\ z_i\end{smallmatrix}\right]^\top P\left[\begin{smallmatrix}1\\ z_i\end{smallmatrix}\right],\ P\succeq 0\right\}. \quad (12)$$

However, the SDP solver SDPT3 cannot solve problems with log determinant terms in the constraints. We will instead create a dummy variable v and solve,

$$P^*{=}\arg\max_{v\le 0, P\in\mathbb{S}^+}\left\{v{+}\log|P| : v \le{-}\left[\begin{smallmatrix}1\\ z_i\end{smallmatrix}\right]^\top P\left[\begin{smallmatrix}1\\ z_i\end{smallmatrix}\right],\ \forall i = 1,\dots m\right\}, \quad (13)$$

where $t^* = v^* + \log|P^*|$ and, thus, is equivalent to Optimization Problem (12).

This problem may be further simplified if we define $z_i(j)$ be the j'th element of the vector $[1, z_i]^\top \in \mathbb{R}^{q+1}$. Then the optimization problem becomes

$$P^* = \arg\max_{v\le 0, P\in\mathbb{S}^+} \ v + \log|P|, \quad (14)$$

$$\text{such that } v \le -\sum_{j,k=1}^{q+1}\sum_{i=1}^{m} z_i(j)z_i(k)P_{j,k}\ \forall\ i,\ P\succeq 0,$$

which is the combination of a $\log\det P$ term and a scalar value, v with m constraints consisting of a linear combination of the elements of P. Problems with the log determinant of a positive semi-definite matrix P, and constraints that consist of linear combinations of elements in P, are convex with respect to the decision variable P [6] and may be solved efficiently with a suitable semi-definite optimization solver such as SDPT3 [7]. Note that, in contrast to the max log likelihood formulation, no analytical solution to the worst-case log likelihood formulation is known.

Example 2: We consider a data set $\mathcal{D} = \{\delta^{(i)}\}_{i=1}^m$, where $m = 600$, generated by a bivariate Gaussian mixture model consisting of 3 weighted Gaussian PDFs. Figures 2(a) and 2(b) display the data set $\mathcal{D}$. Figure 2(a) also shows data enclosing sets, where P^* is the matrix which maximizes the log likelihood in Optimization Problem (13). Figure 2(b) shows the data enclosing sets of the form

$$S(\beta) = \left\{\delta\in\mathbb{R}^n : \left[\begin{smallmatrix}1\\ z_d(\delta)\end{smallmatrix}\right]^\top P^*\left[\begin{smallmatrix}1\\ z_d(\delta)\end{smallmatrix}\right] \le \beta\right\},$$

where P^* is the matrix which maximizes the log likelihood in Optimization Problem (8). Fig. 2(b) displays the same sets but where P^* is obtained from Optimization Problem (13). In both cases β is selected to contain 66% or 99% of the data. Note that (13) yields a tighter set than (8). Tighter data-enclosing sets enable reducing the conservatism in the model by eliminating regions where future data is unlikely to fall. Overly tight sets however, might lack the desired generalization properties we want the solution to have when future data occurs. The numerical complexity of the worst-case log likelihood is studied next.

A. Computational Complexity

Optimization Problem (13) has the same number of variables as Optimization Problem (8), but has an additional m linear constraints. This yields a moderate increase in computation time, still suitable for moderately sized problems with thousands of data points. In Figure 3 we plot the average time taken for a problem with 3 variates and 1000 points for varying monomial degree basis. Both optimization problems

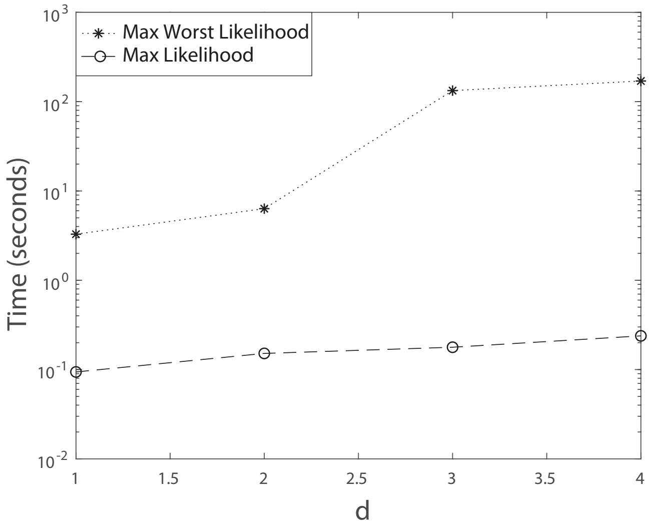

Fig. 3. Average time taken to find P^* for either the max worse likelihood problem or the max likelihood optimization problem for 10 trials.

finish in just a few seconds, but we do see that the additional linear constraints cause (13) to take longer than (8).

In the numerical results section we will see that the volume of sets generated by (13) is often significantly smaller than that for sets based on (8). The increase in computational complexity required by the worst-case likelihood formulation may therefore be justifiable by the corresponding decrease in the conservatism of the set.

V. Parameter Dependencies

In this section we propose a method for relaxing the degree of dependence among the random variables of a Sliced-Normal distribution. Furthermore, we propose a means to penalize solutions that are optimal for the given data, $\mathcal{D}$, but perform poorly on future data - a phenomenon oftentimes called overfitting.

To that end, we will modify the Optimization Problem in (8) by adding a L_1-norm constraint on the off-diagonal elements of P:

$$\max_{P\in\mathbb{S}^+}\left\{m\log|P|-\sum_{i=1}^{m}V_i^TPV_i : P\succeq 0,\ \sum_{i\neq j}^{q}|P_{i,j}|\leq\epsilon\right\}, \tag{15}$$

where ϵ is a fixed constant prescribed in advance. By making ϵ small we obtain a more sparse P, thereby eliminating weak parameter interactions in the SOS. Some such interactions might be the result of having independent parameters. The rationale for evaluating the dependency between any pair of parameters in δ is introduced next.

The key argument of the exponential of a Sliced Normal is the sum of squares $\phi(\delta;P)=(z(\delta)-\mu)^\top P(z(\delta)-\mu)$, which can be written as

$$\phi(\delta;P)=\phi(\delta;U)+\phi(\delta;V)+\phi(\delta;W), \tag{16}$$

where $\phi(\delta;U)$ is not a function of a subset of δ, denoted δ_ℓ, $\phi(\delta;V)$ is not a function of the remaining parameters δ_m, and $P=U+V+W$ is a matrix decomposition satisfying $U_{i,j}V_{i,j}=0$, $U_{i,j}W_{i,j}=0$ and $V_{i,j}W_{i,j}=0$ for all components of P. If $W=0$ we can write the corresponding n-variate Sliced Normal density as the product of two Sliced Normal densities. Because one of them does not depend on δ_ℓ, and the other one does not depend on δ_m, the parameters δ_ℓ and δ_m of the sliced-normal are, therefore, independent. For a fixed P, a measure of the level of dependency between δ_ℓ and δ_m is given by $\|W\|_1$. As expected, there exists a different matrix decomposition for each pair of parameters in δ.

In the developments that follow, we solve Optimization Problem (15) for a fixed value of ϵ, and use the above developments to determine the degree of dependence among all pairs of variables in δ. Assuming $\epsilon=\infty$ allows free dependency modeling, whereas $\epsilon=0$ yields a model in which all the parameters in z-space are independent (not the notion of independence in δ-space we are interested in). The L_1 constraint enables eliminating weak parameter interactions caused by parameter dependencies and outliers. The resulting sliced-normal is an acceptable uncertainty model when the consideration of the constraint does not significantly lower the likelihood of the data (so the dependence between the chosen pair of parameters is weak), and the value of this likelihood is high (so a sliced-normal is a good estimator of the observations). By studying the dependence of the likelihood of the data on ϵ, and progressively decreasing the value of ϵ, weak and spurious interactions/dependencies are systematically identified and eliminated from the model.

To avoid overfitting the training data in $\mathcal{D}$, the value of ϵ will be set according to the log likelihood of the test data $\tilde{\mathcal{D}}$. By finding a value of ϵ that maximizes the log likelihood of a test set of data, we aim to find a distribution that will perform well on future data. A framework for this analysis is presented next.

Let $\tilde{\mathcal{D}}_\delta$ and $\mathcal{D}_\delta$ be two sufficiently large data sequences both drawn from the same probability distribution function. To determine a suitable value for ϵ perform the following steps:

1) Select a minimum, ϵ_L and a maximum value, ϵ_H for the ϵ parameter as well as an increment Δ.
2) Generate a vector v of length n_v that spans ϵ_L to ϵ_H with increment Δ.
3) Use the data points in $\mathcal{D}$ to solve for P_i^* in Optimization Problem (15) using $\epsilon=v_i$.
4) For each P_i^* matrix determine the log likelihood, L_i, of the data in the sequence $\tilde{\mathcal{D}}$.
5) Choose $\epsilon=v_{i^*}$ where $i^*=\arg\max_i L_i$.

Figure 4 shows the likelihood of the test set of data for varying values of ϵ for a data set, $\mathcal{D}$, with $n=7$ parameters along with the optimal value of the objective function for Optimization Problem (15)[1].

[1] The underlying probability density function that generated $\mathcal{D}$ and $\tilde{\mathcal{D}}$ is defined by 7 parameters, δ_i. Of these, $(\delta_1,\delta_2)\sim\mathcal{N}(\Sigma,\mu)$ for some Σ,μ. Meanwhile, $\delta_3=p(\delta_1,\delta_2)$ where $p\in\mathbb{R}(\delta)$ is a randomly generated third degree polynomial function. Next $(\delta_4,\delta_5)\sim\mathcal{N}(\Sigma,\mu)$ for some other Σ,μ. Finally, δ_6, and δ_7 are defined by a PDF uniformly distributed over a circular ring.

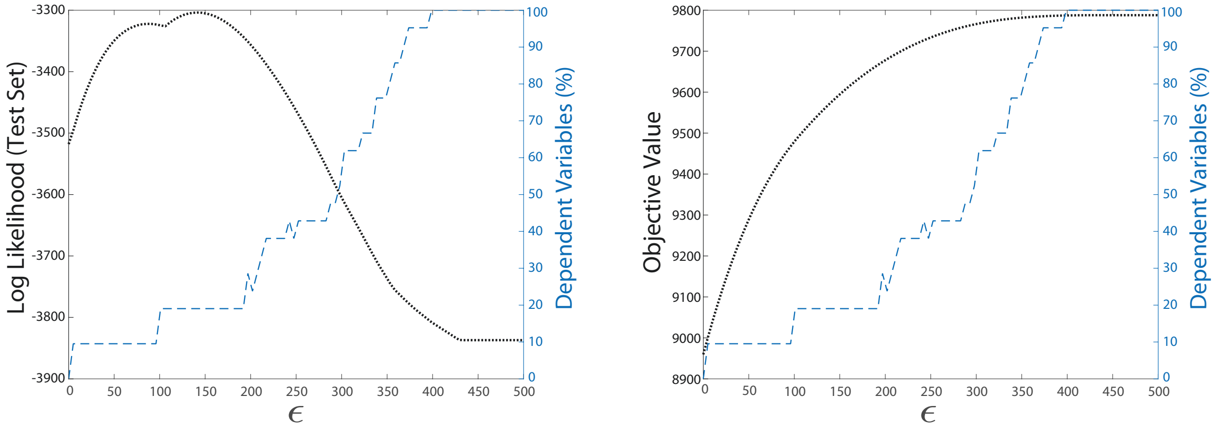

Fig. 4. The subplot on the left plots the log likelihood of the test data for a discrete set of ϵ values. The subplot on the right plots the objective value of the optimization problem from which P^* is calculated for the same set of ϵ values. The blue dashed-line is the percent of dependent variables in the resulting model with respect to the value of ϵ and is overlayed on both subplots.

Based on Fig. 4 we select $\epsilon = 141.41$ for this data set. Note that once ϵ is larger than the L_1 norm of the unconstrained optimal matrix P^*, the L_1 constraint no longer effects the objective value of Optimization Problem (15). We see that the objective function is, as one would expect, maximized as fewer terms are forced to be independent, but that the function performs best on the test set when only approximately 20% of the terms are dependent.

For estimating whether two variables are dependent in Fig. 4, we say that if any element of P^* corresponding to monomials of the parameters δ_l, and δ_k is greater than 10^{-6}, then the parameters are dependent. If no such element of P^* exists, then δ_l and δ_k are independent. We calculate the percentage of dependent parameters as being equal to the number of unique dependent parameters, divided by the number of unique parameter pairings. In the $n = 7$ case there are 21 unique combinations of two variables, and in this particular example 5 of these unique combinations are, in fact, dependent. Using the matrix P^* derived from (15) for $\epsilon = 141.41$, we are correctly able to determine which of the parameters are independent without a priori knowledge on the probability distribution function that generated the data.

VI. Numerical Experiments

In this section we illustrate the ideas above by finding semi-algebraic sets that tightly enclose the data, and by seeking uncertainty models for which weak parameter dependencies are eliminated.

A. *Data-enclosing Sets*

Next we generate data-enclosing sets given data sequences drawn from several data-generating processes. The first group of sets, called 11-Lin, has 100 data sequences, $\mathcal{D}$, drawn from an 11-dimensional multivariate normal distribution having a random number of dependent parameters. One of such sequences is shown in Figure 5. The second group, called 4-Non, has 100 data sequences drawn from a 4-dimensional random vector for which the first two parameters are normally distributed whereas the other two are uniformly distributed over a circular ring. Finally the last data set, called 2-Non, has 50 data sequences drawn from a 2-dimensional vector uniformly distributed over a circular ring. Each data sequence has $m = 500$ data points.

For each data set we find the optimal density function using Optimization Problem (8) and (13) and the tightest data-enclosing set $S(\beta^*)$ where,

$$\beta^* = \min_{\delta \in \mathcal{D}} (Z_d(\delta) - \mu)^\top P (Z_d(\delta) - \mu),$$

for Optimization Problem (8) and

$$\beta^* = \min_{\delta \in \mathcal{D}} \left(\begin{bmatrix} 1 \\ Z_d(\delta) \end{bmatrix}^\top P \begin{bmatrix} 1 \\ Z_d(\delta) \end{bmatrix} \right),$$

for Optimization Problem (13).

We then generate a test set of 2000 data points from the same probability distribution function and evaluate the fraction of such points contained by S. In addition, we generate 500000 uniformly distributed points over the smallest hyper-rectangle containing the test points. By evaluating the percentage of these points which fall within S we have an approximate metric proportional to the volume of the set. Table I summarizes the results. There we see that for the Gaussian multivariate data, a degree 2 polynomial set performs well in both cases. The maximum log likelihood formulation captures a larger number of test points, but its area is almost 8 times larger than the worst case likelihood formulation.

For the 4-Non dataset, the max log likelihood approach of degree 4 had a larger approximate area and a smaller percentage of correctly labeled test data than that of degree 2. The optimal worst-case log-likelihood function, however, generated a semi-algebraic set which had significantly less area, almost 4 times less, but also had a lower effectiveness of predicting future points of 92.61%.

Finally, the 2-Non dataset demonstrated that a higher degree polynomial can decrease the area of the set while still retaining the same predictive capability. We see that the maximum log likelihood formulation of degree 4 led to sets that practically contained the same percentage of test points

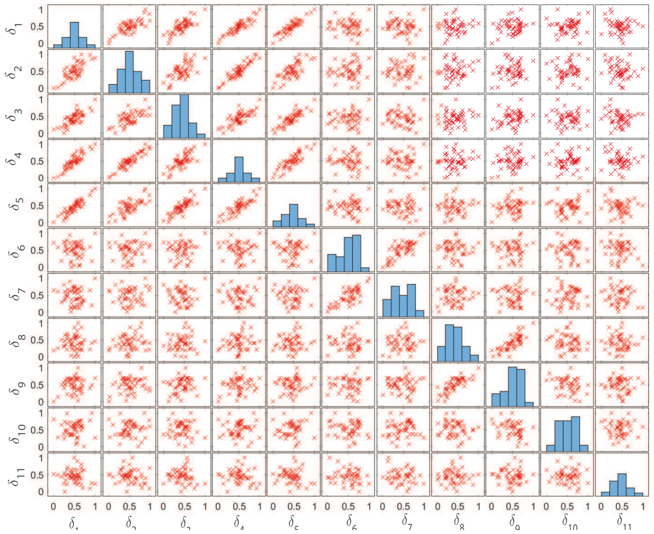

Fig. 5. Normalized data generated by an 11 variate Gaussian dataset used for testing the accuracy of the parameter dependence method. Plots along the diagonal of the figure are histograms of a parameter, off diagonal plots are the scatterplot of two parameters. Here the first five variates are linearly dependent on each other, the sixth and seventh variates are linearly dependent and the eighth and ninth variates are linearly dependent. The tenth and eleventh variate are independent of all other variates.

than that for the degree 2, but its volume estimate was almost half that of the degree 2. Increasing the degree in this case decreased the conservativeness of the set estimate.

TABLE I

ACCURACY OF THE PROPOSED METHOD FOR FINDING SEMI-ALGEBRAIC SETS WITH HIGH LIKELIHOOD OF CONTAINING FUTURE POINTS. TESTS WERE PERFORMED OVER A TEST SET OF 2000 DATA POINTS FOR A MULTIVARIATE GAUSSIAN DATA SET AND A NUMBER OF DATA SETS WITH NONLINEAR DEPENDENCIES.

Data Set	Method	Degree	Test Set Correct	Vol. Est.
11-Lin	Max log like	2	99.58 %	0.39 %
11-Lin	Worst-case like	2	90.04 %	0.05 %
4-Non	Max log like	2	99.79 %	44.25 %
4-Non	Worst-case like	2	97.90 %	34.21 %
4-Non	Max log like	4	99.69 %	57.41 %
4-Non	Worst-case like	4	92.61 %	10.39 %
2-Non	Max log like	2	99.74 %	84.76 %
2-Non	Worst-case like	2	99.29 %	80.23 %
2-Non	Max log like	4	99.79 %	48.73 %
2-Non	Worst-case like	4	97.63 %	34.31 %

B. Identifying Parameter Dependencies

Next we study the dependency among the parameters by using the step-by-step procedure of Section V with Optimization Problem (15) and $d = 1$. When $d = 1$ we are fitting a Gaussian in the original parameter space, and every off-diagonal term of P models a dependency between parameters. Two groups of data sets were considered. For the first group we draw 100 data sequences from a multivariate Gaussian distribution of 11 variables. In each data sequence, k, the number of groups of independent variables are randomly varied between 2 and 8. We use only $m = 200$ points in the training data set and 200 points in the testing data set. The second group is comprised of 50 data sequences drawn from a 7-variate random vector exhibiting non-Gaussian dependencies. The groups of independent variables are $1, 2, 3, 4, 5$; and $6, 7$. We use 500 training and 500 testing points for this case.

If we consider every unique pairing of parameters, then Table II shows the percentage of independent parameters that were correctly identified as being independent (True Positive) as well as the percentage of dependent parameters that were correctly identified as being dependent (True Negative). In both cases, the proposed approach correctly identifies a majority of the dominant dependencies.

TABLE II

ACCURACY OF THE PROPOSED METHOD FOR DETERMINING INDEPENDENCE OF VARIABLES, 11-LIN IS A MULTIVARIATE GAUSSIAN DATA SET WITH 11 VARIABLES AND 7-NON IS A 7 VARIABLE DATA SET WITH NONLINEAR DEPENDENCE BETWEEN VARIABLES.

Data Set	True Positive	True Negative
11-Lin	99.66 %	77.41 %
7-Non	75.20%	96.75 %

VII. CONCLUSION

This paper proposes SOS optimization techniques for creating computational models of multivariate data. The proposed techniques can be used to model multimodal phenomena exhibiting strong parameter dependencies as well as to generate semi-algebraic sets that tightly enclose the data. Furthermore, we present an approach that can be used to quantify dependencies between variables using the data, so that weak and spurious dependencies resulting from small data sets or outliers can be systematically eliminated. The proposed approach is based on identifying an optimal Gaussian density in a higher-dimensional feature space, and projecting a polynomial slice of the resulting density onto physical space. As such, the resulting distribution is called a Sliced Normal. Sliced Normals are a versatile family of random variables which naturally characterize complex parameter dependencies. The characterization of such dependencies, usually omitted in practice, is instrumental in reducing the considerable conservatism incurred in standard practices in uncertainty quantification, system identification, robust control analysis, and robust control design.

REFERENCES

[1] L. Crespo, G. D., K. S., and D. J., “A scenario optimization approach to system identification with reliability guarantees,” *American Control Conference*, 2019.

[2] H. El-Samad, S. Prajna, A. Papachristodoulou, M. Khammash, and J. Doyle, “Model validation and robust stability analysis of the bacterial heat shock response using sostools,” 2003.

[3] G. Chesi, “Lmi techniques for optimization over polynomials in control: a survey,” *IEEE Transactions on Automatic Control*, 2010.

[4] I. Myung, “Tutorial on maximum likelihood estimation,” *Journal of mathematical Psychology*, 2003.

[5] L. Gu, “Multivariate gaussian distribution,” CMU, Tech. Rep., 2008.

[6] K. Toh, M. Todd, and R. Tütüncü, “On the implementation and usage of sdpt3–a matlab software package for semidefinite-quadratic-linear programming, version 4.0,” in *Handbook on semidefinite, conic and polynomial optimization*, 2012.

[7] K. Toh, M. Todd, and R. Tutuncu, “Sdpt3 - a matlab software package for semidefinite programming,” *Optimization Methods and Software*, 1999.