Dimensionality Reduction of Sliced-Normal Distributions

Luis G. Crespo * **Brendon K. Colbert** * **Sean P. Kenny** * **Daniel P. Giesy** *

* *Dynamic Systems & Control Branch, NASA Langley Research Center, Hampton, VA, 23681, USA. e-mail: luis.g.crespo@nasa.gov.*

Abstract: Sliced-Normal (SN) distributions enable the characterization of complex multivariate data as both a vector of possibly dependent random variables and as a semi-algebraic, tightly enclosing set. SNs inject the physical space into a higher dimensional (so-called) feature space using a polynomial mapping. Optimization-based strategies for estimating SNs from data in both physical and feature space were recently developed. The formulations in physical space yield non-convex optimization programs whose solutions exhibit the best performance, whereas the formulations in feature space yield either an analytical solution or a convex program thereby facilitating their application to higher dimensional datasets. In both cases, however, the exponential dependency of the number of optimization variables on the dimension of feature space limits their application to moderately sized problems. Two strategies to mend for this deficiency are proposed herein. The first strategy identifies groups of highly interdependent parameters exhibiting a possibly nonlinear dependency using a distribution-free framework. This classification enables estimating a SN for any of such groups independently of the other groups thereby reducing the computational complexity of the estimation process. The second strategy reduces the dimension of feature space by only retaining the monomials of the polynomial mapping that significantly increase the likelihood of the data while leveraging lower dimensional SNs. A system identification example is used for illustration.

Keywords: Uncertainty quantification, sliced-normals, system identification, dependency analysis.

1. INTRODUCTION

The characterization of the distribution of data as well as of their spread is of great significance to system identification and robust control. Variability in measured data arises from aleatory variation in physical parameters, varying operating conditions, model-form uncertainty, and measurement error. In contrast, variability in calculated data arises from numerical, approximation and convergence errors. This characterization is particularly challenging when the data exhibits strong parameter dependencies. Such dependencies are often modeled using copulas (Nelsen (2007); Kurowicka and Cooke (2006)). Copulas are estimated by determining an individual marginal distribution for each variable separately, and then adding a dependence structure that captures the joint behavior. In the multivariate case, this behavior is often modeled by the pair-copula decomposition approach introduced as C-vine and R-vine copulas (Kjersti et al. (2009); Czado (2019); Haff et al. (2013)). Several parametric and non-parametric families of copula models can be used to estimate dependencies. The acceptability of a dependency model depends strongly on the selection of an appropriate family of copulas. Unfortunately, standard families of copulas often fail to accurately describe complex dependencies.

A Sum of Squares (SOS) optimization approach to modeling multivariate data using SNs was recently proposed in Crespo et al. (2019a). A system identification strategy based on SNs is outlined therein, whereas Crespo et al. (2019b) exemplifies its application. In contrast to copulas, SNs model the marginal distributions of individual parameters and their joint behavior simultaneously. Numerical experiments suggest that SNs are more versatile than most copula families since they can handle multi-modal distributions with non-monotonic dependencies. Furthermore, whereas the selection of a copula structure is a cumbersome process requiring extensive expertise, the selection of a SN structure only requires prescribing the degree of a polynomial. However, as with all other SOS methods, the computational cost of estimating a SN restricts its application to moderately sized problems [1]. This paper develops strategies for extending the SN's range of application by lowering this cost.

2. SLICED-NORMAL DISTRIBUTIONS

This section summarizes key developments of Crespo et al. (2019a). The Normal density [2] of $z \in \mathbb{R}^{n_z}$ is given by

$$f_z(z; \mu, P) = \frac{1}{\nu} e^{\frac{-\phi(z,\mu,P)}{2}}, \tag{1}$$

where $\mu \in \mathbb{R}^{n_z}$ is the mean, $P \in \mathbb{R}^{n_z \times n_z}$ is the inverse of the covariance matrix, $\nu = (2\pi)^{\frac{n_z}{2}} \sqrt{\det(P^{-1})}$, and

$$\phi(z, \mu, P) = (z - \mu)^\top P(z - \mu). \tag{2}$$

P is symmetric and positive definite, which will be denoted as $P \succ 0$. Consider the polynomial mapping from physical space δ to feature space z given by the function

$$z = Z(\delta, d), \tag{3}$$

where $Z(\delta, d) : \mathbb{R}^{n_\delta} \to \mathbb{R}^{n_z}$ with $n_z = \binom{n_\delta+d}{n_\delta} - 1$, is the vector of monomials in variables δ of degree greater than zero and less than or equal to d. The monomials of $Z(\delta, d)$ will be in graded lexicographic order: they are first ordered by the canonical order in the degree, and, second by using lexicographic order based

[1] SNs that maximize the likelihood of the data in feature space are only restricted by the ability to invert a matrix. As such, they are readily applicable to large dimensional datasets and high-dimensional feature spaces.

[2] The density and the Cumulative Distribution Function (CDF) of z with parameter θ will be denoted as $f_z(z; \theta)$ and $F_z(z; \theta)$ respectively.

on $\delta_1 = a$, $\delta_2 = b$, ... For example, if $n_\delta = 2$, then $Z(\delta, 2) = [\delta_1, \delta_2, \delta_1^2, \delta_1\delta_2, \delta_2^2]^\top$.

The joint density of a *Sliced-Normal* distribution supported in $\Delta \subset \mathbb{R}^{n_\delta}$ is defined as

$$f_\delta(\delta; \mu, P, d, \Delta) = \begin{cases} \frac{\nu}{c} f_z(Z(\delta, d); \mu, P) & \text{if } \delta \in \Delta, \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where d, μ, P and Δ are the parameters of the distribution and

$$c(d, \mu, P, \Delta) = \int_\Delta \nu f_z(Z(\delta, d); \mu, P)\ d\delta, \tag{5}$$

is a normalization constant (see Section 3 for a method for estimating c). Hence, the density of a SN results from evaluating the Normal density (1) on the polynomial manifold given by $Z(\delta, d)$ for all $\delta \in \Delta$, and performing the corresponding normalization. This gives rise to the SN name. Note that the minus argument of the exponential in f_δ, $\phi(Z(\delta, d), \mu, P)$, is a SOS of polynomials in δ. Features of the physical variables, such as boundedness or positiveness, can be enforced by choosing a suitable Δ. In the following developments the SN parameters d and Δ will be fixed upfront. Therefore they will be dropped from the notation when their role is not essential.

The superlevel sets of f_δ are

$$S\left(\hat{\phi}, \mu, P\right) = \left\{\delta \in \Delta : \phi(z, \mu, P) \leq \hat{\phi}\right\}, \tag{6}$$

where $\hat{\phi}$ is a fixed constant. Equation (6) defines a nested family of closed, semi-algebraic sets satisfying $S(\phi_1, \mu, P) \subseteq S(\phi_2, \mu, P)$ for $\phi_1 < \phi_2$. The polynomial structure of these sets makes them suitable for rigorous worst-case analysis and design methods.

2.1 *Estimation of Sliced Normals*

Assume that n independent and identically distributed observations of a stationary Data Generating Mechanism are available. Denote the physical parameter as $\delta \in \mathbb{R}^{n_\delta}$, and the corresponding data sequence as $\mathcal{D} = \{\delta^{(1)}, \ldots, \delta^{(n)}\}$. Means of estimating μ and P according to the data sequence $\mathcal{D}$ using Maximum Likelihood (ML) formulations are presented next.

Optimality in Physical Space: The SN that maximizes the (log) likelihood of the data in physical space,

$$\mathcal{L}(\mu, P, \mathcal{D}) = \sum_{i=1}^{n} L_\delta\left(\delta^{(i)}, \mu, P\right), \tag{7}$$

where $L_\delta(\delta, \mu, P) = \log f_\delta(\delta; \mu, P)$, is given by

$$\langle \mu^\star, P^\star \rangle = \underset{\mu \in \mathbb{R}^{n_z},\ P \succ 0}{\text{argmax}} \left\{ -n \log c - \frac{1}{2} \sum_{i=1}^{n} \phi\left(z^{(i)}, \mu, P\right) \right\}, \tag{8}$$

where $z^{(i)} \triangleq Z\left(\delta^{(i)}, d\right)$ for $i = 1, \ldots n$. The integration constant c makes (8) a non-convex optimization program. This program, to be referred to as ML$_\delta$ hereafter, can be solved using standard gradient-based algorithms.

Optimality in Feature Space: The SN that maximizes the (log) likelihood of the data in feature space,

$$\mathcal{L}\left(\mu, P, Z(\mathcal{D}, d)\right) = \sum_{i=1}^{n} L_z\left(z^{(i)}; \mu, P\right), \tag{9}$$

where $L_z(z; \mu, P) = \log f_z(z; \mu, P)$, is given by

$$\langle \mu^{\star}, P^{\star} \rangle = \underset{\mu \in \mathbb{R}^{n_z},\ P \succ 0}{\text{argmax}} \sum_{i=1}^{n} L_z\left(z^{(i)}; \mu, P\right), \tag{10}$$

whose solution is

$$\mu^{\star} = \frac{1}{n} \sum_{i=1}^{n} z^{(i)}, \tag{11}$$

$$P^{\star} = \left(\frac{1}{n} \sum_{i=1}^{n} \left(z^{(i)} - \mu^{\star}\right) \left(z^{(i)} - \mu^{\star}\right)^\top \right)^{-1}. \tag{12}$$

Equations (11) and (12) are the empirical mean and the inverse of K, the empirical covariance of the data in feature space. This formulation will be denoted as ML$_z$ hereafter.

A variant of ML$_z$ that improves $\mathcal{L}(\mathcal{D})$ by augmenting the covariance of the SN (Colbert et al. (2020)) uses $\mu = \mu^{\star}$ and $P = \gamma^{\star} P^{\star}$, where $\gamma^{\star}$ is the solution to

$$\max_{\gamma \in \mathbb{R}^+} \left\{ -2n \log c\left(\mu^{\star}, \gamma P^{\star}\right) - \sum_{i=1}^{n} \phi\left(z^{(i)}, \mu^{\star}, \gamma P^{\star}\right) \right\}. \tag{13}$$

This convex program will be denoted as ML$_z^+$ hereafter. Note that the collection of superlevel sets in (6) for ML$_z$ and ML$_z^+$ are the same. The formulation in physical space often yields SNs that model the distribution of the data better than those in feature space. However, ML$_z$ and ML$_z^+$ allow efficient modeling of data sequences with large n_z's.

Example 1: The versatility of the SNs is illustrated by considering the Van der Pol oscillator. In this example δ is the state and the initial condition is an uncorrelated normal with means and standard deviations equal to $1/4$ and $1/10$ respectively. Figure 1 shows the data sequence corresponding to the time response in the $[0, 7]s$ time interval, as well as the density of the estimated ML$_\delta$ and ML$_z$ SNs for $d = 4$. The peaks/modes of the distribution near the limit cycle correspond to states where the dynamics are slower.

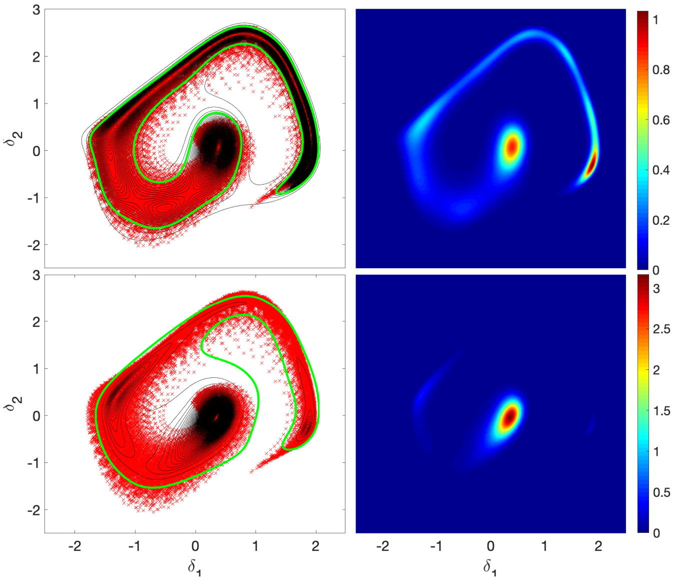

Fig. 1. ML$_\delta$ SN (top) and ML$_z$ SN (bottom). The left subplots show observations (×), the boundaries of the sets S in (6) (black lines), and the set containing 95% of the data (green line), whereas the right subplots show the joint density.

3. COMPUTATIONAL COMPLEXITY

Large values of n_δ and d complicate the numerical search for the optimum of (8) and the calculation of the inverse of the positive definite matrix in (12). The rapid increase in the dimension of the feature space n_z that results from increasing values of n_δ and d is illustrated in Table 1. The entries of this table prescribe the size of the matrix in (12) to be inverted. Numerical experiments show inaccurate inverse matrices and long computational times for n_z exceeding 5000 (assuming P is not nearly singular).

Table 1. Dimension of feature space n_z.

	$d=2$	$d=3$	$d=4$	$d=5$
$n_\delta=2$	5	9	14	20
$n_\delta=5$	20	55	125	251
$n_\delta=10$	65	285	1000	3002
$n_\delta=15$	135	815	3875	15503

Seeking a SN using ML_δ requires the repeated estimation of the integration constant in (5). The number of such calculations grows with the number of decision variables of (8), which are listed in Table 2. In addition to the large number of required estimations of c, the value of n_δ also affects the accuracy of the estimates. A sampling-based approximation to c is

Table 2. Number of decision variables of (8).

	$d=2$	$d=3$	$d=4$	$d=5$
$n_\delta=2$	20	54	119	230
$n_\delta=5$	230	1595	8000	31877
$n_\delta=10$	2210	41040	501500	4510505
$n_\delta=15$	9315	333335	7513625	120194759

$$c(\mu, P) = \frac{\text{Vol}\,[\Delta]}{m} \sum_{i=1}^{m} f_\delta\left(\delta_u^{(i)}; d, \mu, P, \Delta\right), \tag{14}$$

where $\delta_u^{(i)}$ for $i = 1, \dots m$ are samples uniformly distributed over Δ and $\text{Vol}\,[\cdot]$ is the volume operator. Equation (14) allows using the same set of samples during the search for $\langle \mu^\star, P^\star \rangle$. However, the approximation becomes inaccurate when $f_\delta(\delta_u^{(i)}; d, \mu, P, \Delta) \approx 0$ for most of the m samples. This will likely be the case when m is too small for the corresponding n_δ. Methods that perform integration by quadrature suffer from similar deficiencies in large dimensions. Furthermore, large values of n_δ complicate the sampling of a SN regardless of the formulation used to estimate it. Even though Markov Chain Monte Carlo methods do not require evaluating c to sample a SN, the chance of the random walk missing isolated modes or producing repeated samples increases with this dimension.

A means to lower the complexity of the above tasks is to use dependency information to group the n_δ parameters according to their degree of interdependency. By estimating a SN for each group of strongly interdependent parameters independently of the rest, we replace a single larger n_δ (and, thus, n_z) by several smaller ones. This practice also enables adjusting the fidelity of the SN for each group of parameters according to the complexity of the particular parameter dependency. This is the rationale of the first model reduction technique presented below. Another approach is to eliminate the monomials in the polynomial mapping that do not significantly increase the likelihood of the data. This can be accomplished by starting from a low-dimensional SN and then testing whether an additional monomial improves the model significantly or not. The monomials that do so will be kept while the rest will be discarded. This is the rationale of the second model reduction technique proposed.

4. DEPENDENCY ANALYSIS

Denote as $g_1 = \{\delta_1, \dots \delta_{n_1}\}$ and $g_2 = \{\delta_{n_1+1}, \dots \delta_{n_\delta}\}$ two groups of parameters having no elements in common and whose union is equal to δ. When all of the elements of g_1 are independent from all of the elements of g_2 we have

$$f_\delta = f_{g_1} f_{g_2}. \tag{15}$$

Means to model g_1 and g_2 as independent multivariate SNs are considered next. Assume that $Z(\delta, d)$ has been rearranged as

$$Z(\delta, d) = \left[m_1(g_1),\ m_2(g_2),\ Z(\delta, d) \setminus (m_1 \cup m_2)\right], \tag{16}$$

where m_1 contains all the monomials in $Z(\delta, d)$ that only depend on the elements of g_1, and m_2 contains all the monomials in $Z(\delta, d)$ that only depend on the elements of g_2. For the independence condition (15) to hold, it is required that

$$P = \begin{bmatrix} P_1 & 0 & 0 \\ 0 & P_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{17}$$

where $P_1 \succ 0 \in \mathbb{R}^{a \times a}$ with $a = \binom{n_1+d}{n_1} - 1$, and $P_2 \succ 0 \in \mathbb{R}^{b \times b}$ with $b = \binom{n_\delta - n_1 + d}{n_\delta - n_1} - 1$. The monomials combining elements of g_1 and g_2 introduce the zero rows and columns in P. When the parameters in g_1 are weakly dependent on the parameters in g_2, we can estimate a SN for the elements of g_1 independently from the SN for the elements of g_2. This practice enables the analyst to solve two smaller optimization programs: one in which $Z(\delta, d) = m_1(g_1)$ and the unknowns are μ_1 and P_1, an another one where $Z(\delta, d) = m_2(g_2)$ and the unknowns are μ_2 and P_2. This rationale can be extended to a multi-group setting where $\delta = \cup_i^{n_g} g_i$ and $g_i \cap g_j = \emptyset$ for $i \neq j$. Note that the pursuit of a SN comprised of independent groups of parameters by enforcing sparsity conditions to P will suffer from the numerical difficulties explained above. Furthermore, the resulting P becomes singular as the numerical search approaches the optimum regardless of the value of n_z. Means to identify the non-overlapping groups $g_1, g_2, \dots, g_{n_g}$ are presented next.

4.1 Parameter Grouping

A means to evaluate the degree of dependency among the components of δ seen in $\mathcal{D}$ is presented first. The copula of the continuous random vector δ with joint CDF F_δ is defined as the joint CDF of u, where $u = F_\delta(\delta)$, i.e.,

$$F_u(u) = \mathbb{P}\left[\delta_1 \leq F_{\delta_1}^{-1}(u_1), \dots, \delta_{n_\delta} \leq F_{\delta_{n_\delta}}^{-1}(u_{n_\delta})\right]. \tag{18}$$

A copula contains all information on the dependence structure between the components of δ independently of its marginals. When the components of δ are independent (18) reduces to

$$F_u^{\text{ind}}(u) = \prod_{i=1}^{n_\delta} u_i. \tag{19}$$

When F_δ is unknown but $\mathcal{D}$ is available, we can approximate (18) with the empirical copula

$$F_u^{\text{emp}}(u; \mathcal{D}) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}\left(F_\delta\left(\delta^{(j)}; \mathcal{D}\right) \leq u\right), \tag{20}$$

where $\mathbf{1}(\cdot)$ is the indicator function, and

$$F_{\delta_i}(\delta_i; \mathcal{D}) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}\left(\delta_i^{(j)} \leq \delta_i\right), \tag{21}$$

the empirical CDF of δ_i, is a component of $F_\delta(\delta, \mathcal{D})$.

The degree of dependency among the components of δ will be evaluated using

$$\rho(\delta_1, \dots, \delta_{n_\delta}, \mathcal{D}) = \mathbb{V}\left[F_u^{\text{ind}}(u) - F_u^{\text{emp}}(u; \mathcal{D})\right], \qquad (22)$$

where $\mathbb{V}[\cdot]$ is the variance operator, and u is a uniform random vector in $[0, 1]^{n_\delta}$. This equation can be readily evaluated by sampling. When the components of δ are weakly dependent $\rho(\delta_1, \dots, \delta_{n_\delta}, \mathcal{D}) \approx 0$. The degree of dependency between the elements of g_1 and g_2 will be evaluated using

$$\rho(g_1, g_2, \mathcal{D}) = \max_{i,\, j} \left\{\rho(\delta_i, \delta_j, \mathcal{D}) \,:\, \delta_i \in g_1,\ \delta_j \in g_2\right\}. \qquad (23)$$

Therefore, the degree of multivariate dependency between g_1 and g_2 is given by the largest degree of bivariate dependency among all possible parameter pairs. This implies that the groups g_1 and g_2 will be considered weakly dependent when all possible pairs of parameters are weakly dependent, i.e., when $\rho(\delta_i, \delta_j, \mathcal{D})$ is below a cut-off value ϵ. Otherwise, g_1 and g_2 will be considered dependent to a degree proportional to ρ. An algorithmic implementation of these ideas is presented next.

Dependency Analysis Algorithm: Assume that the data sequence $\mathcal{D}$ is available and the cut-off value ϵ has been fixed. Instantiate the set of groups as $g_1 \leftarrow \{\delta_1\}$, the number of groups as $n_g \leftarrow 1$, and proceed as follows:

(1) Let $\bar{\delta}$ be any element of $\delta \setminus \cup_{i=1}^{n_g} g_i$.

(2) **For** $i = 1, \dots n_g$ **do** calculate $\rho(g_i, \bar{\delta})$ using (23) **end do**

(3) Determine $s = \{i \subseteq (1, \dots, n_g) \,:\, \rho(g_i, \bar{\delta}) \geq \epsilon\}$.

(4) If $s = \emptyset$, then $g \leftarrow \{g, \bar{\delta}\}$ and $n_g \leftarrow n_g + 1$. Otherwise, in g replace the groups g_i, $i \in s$ by the single group $\cup_{i \in s} g_i \cup \bar{\delta}$ and update n_g accordingly.

(5) If $\delta = \cup_{i=1}^{n_g} g_i$ stop. Otherwise, go to Step 1.

Therefore, every parameter within a group will be strongly dependent to at least one parameter of the same group while being weakly dependent to all the parameters of all other groups. By performing this classification, we can model each group independently of the rest thereby reducing the computational complexity of estimating the SN for the full δ.

The value of ϵ can significantly impact the dependency analysis. Overly small values will let weak and spurious dependencies affect the resulting grouping whereas large values will only capture strong dependencies. As such, expert judgment should be used when prescribing such a value. As expected, high-dimensional data sets will require the consideration of a large number of pairs. Processing of each pair however is a computationally efficient task requiring only algebraic manipulations. Equation (23) might not require evaluating all parameter pairs. Once any of the pairs yields $\rho(\delta_i, \delta_2) \geq \epsilon$, the calculation for the remaining pairs can be avoided. The grouping in g along with all the supporting metrics are naturally displayed using an undirected graph (see e.g., Figure 4).

The results of the parameter grouping method are more general and informative than those resulting from correlation analyses because they also capture nonlinear parameter dependencies. Note that this method can be applied outside the context of SNs.

Example 2: In this example we consider the system identification of an unknown plant given the $n = 2000$ time responses to a doublet input shown in Figure 2. To this end we assume the linear time invariant model given by

$$H(s) = \frac{p_1 s^4 + p_2 s^3 + p_3 s^2 + p_4 s + p_5}{p_6 s^5 + p_7 s^4 + p_8 s^3 + p_9 s^2 + p_{10} s + p_{11}}. \qquad (24)$$

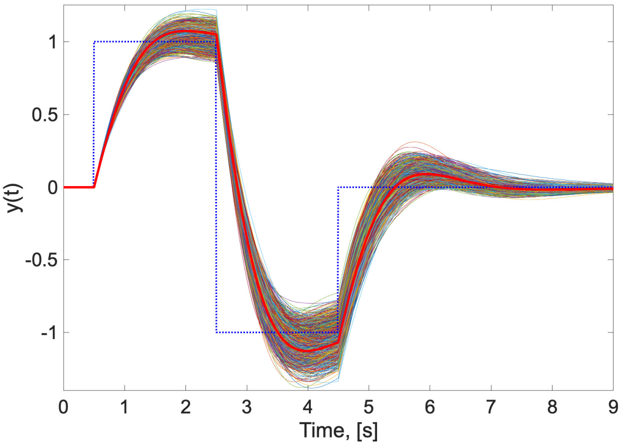

Fig. 2. Time responses to a doublet input. The mean response is shown in red.

The estimation of the parameters of the transfer function for each of the data functions yields the data sequence shown in Figure 3. Each point in the sequence, $\delta^{(i)} \in \mathbb{R}^{11}$, results from minimizing the prediction error. The subplots on the diagonal show the empirical marginal densities whereas the off-diagonal subplots are projections of the cloud of parameters onto 2-dimensional subspaces. Note that the marginals are bell-shaped, and that some significant parameter dependencies exist (dependencies might not be apparent by visually inspecting the spread of data cloud).

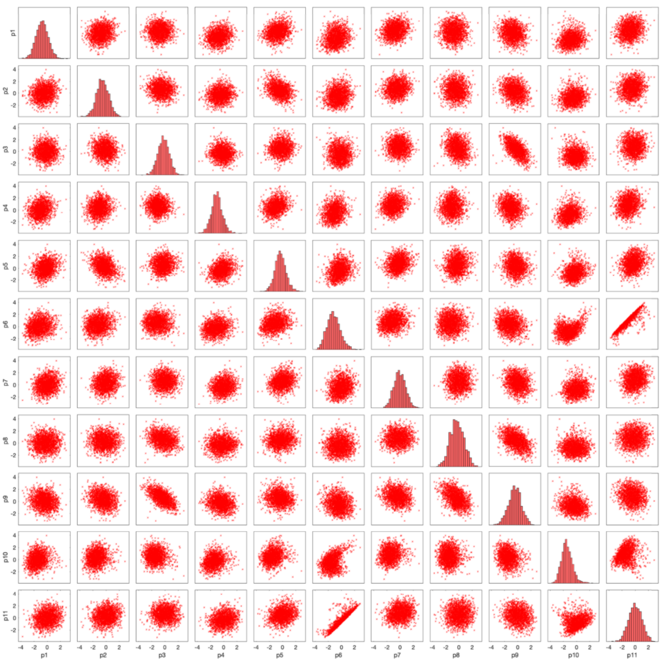

Fig. 3. Data sequence resulting from parameter estimations.

Further notice that the common practice of normalizing the estimated coefficients of the transfer function by the coefficient of the largest power in the denominator, p_6, will induce analyst-made interdependencies. The estimation of a SN using ML_δ for $d = 5$ requires 9541895 decision variables whereas the estimation of a SN via ML_z (or ML_z^+) for the same degree requires inverting a covariance matrix of size 4368 × 4368.

The dependency analysis for $\epsilon = 5$ led to dependency graph in Figure 4. This figure shows the values of $\rho(\delta_i, \delta_j)$ leading to the parameter grouping [8, 3, 9], [1], [2, 5] , [4], [6, 10, 11] and [7]. It is worth noting that this grouping differs from the grouping that results from evaluating the degree of dependency using correlation coefficients. The estimation of SNs of degree 5 for each of these 6 groups can be efficiently performed using any of the ML formulations. Samples of the SNs that result from ML_δ and ML_z^+ (not shown) are practically indistinguishable from those in Figure 3. More importantly, the number of decision variables required by ML_δ are 1595, 20, 230, 20, 1595 and 20 respectively. This is about 0.04% of the number of decision variables of the original program. Furthermore, the proposed approach enables the analyst to adjust the fidelity of the SN for each subgroup, i.e., choosing d according to the complexity of the dependencies of each subgroup. This cannot be accomplished by estimating a SN for the full δ.

In regard to ML_z, the size of the covariances that must be inverted are 55, 5, 20, 5, 55, and 5 respectively. The effort of carrying out such tasks is considerably smaller than inverting the 4368 × 4368 covariance matrix of a fully dependent δ. This illustrates the benefits of accounting for parameter dependencies before estimating a SN model. In addition, estimating SN for each group further increases the likelihood of the data.

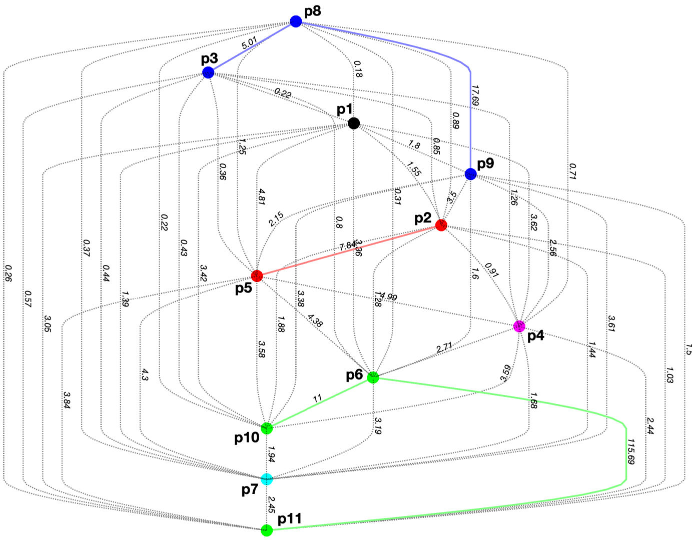

Fig. 4. Dependency graph with supporting $\rho(\delta_i, \delta_j)$ values.

5. MAPPING REDUCTION

The mapping in (3) uses all the monomials of degree greater than zero and less than or equal to d. This practice along with large values of n_δ and d make the estimation of SNs computationally expensive. In this section we propose a sequential approach for estimating SNs that mitigates this cost. This is accomplished by (i) restricting the dimension of feature space so only monomials that significantly increase the likelihood of the data are mapped, and by (ii) using lower-dimensional SNs to calculate higher-dimensional SNs. As before, we will present one approach in physical space and another one in feature space. [3]

5.1 *Maximum Likelihood in Physical Space*

The SN parameter $P \succ 0$ can be written as

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^\top & P_{22} \end{bmatrix}, \tag{25}$$

where $P_{11} \succ 0 \in \mathbb{R}^{p \times p}$, $P_{12} \in \mathbb{R}^{p \times q}$ and $P_{22} \in \mathbb{R}^{q \times q}$. Our goal is to find an optimal P for a given P_{11}. Next we seek to compute the P corresponding to the mapping $Z^+(\delta) = \{Z_1(\delta), Z_2(\delta)\}$, i.e., P in (25), by using the known P corresponding to the mapping $Z = Z_1(\delta)$, i.e., P_{11} in (25). The application of Schur complement properties to P implies that $P \succ 0$ if and only if

$$P_{22} \succ 0 \ \text{ and } P_{11} - P_{12} P_{22}^{-1} P_{12}^\top \succ 0. \tag{26}$$

When $q = 1$ these conditions are equivalent to

$$P_{22} > 0 \text{ and } P_{22} - P_{12}^\top P_{11}^{-1} P_{12} > 0, \tag{27}$$

which is a set of convex constraints in the decision variables P_{12} and P_{22}. Equation (25) and constraints (26) allow for the sequential augmentation of P. This process, which solves for a comparatively small number of decision variables at each step of the sequence, allows for the elimination of terms of the polynomial mapping, i.e., dimensions of future space, that do not increase the likelihood of the data significantly. In particular [4], given a $P_{Z_1}^\star$ corresponding to the baseline mapping $Z_1(\delta)$, the parameters of a SN corresponding to the augmented mapping $Z^+(\delta) = \{Z_1(\delta), Z_2(\delta)\}$ are given by

$$\langle \mu_{Z^+}^\star, P_{12}^\star, P_{22}^\star \rangle = \underset{\mu \in \mathbb{R}^{n_z},\, P_{12},\, P_{22}}{\text{argmax}} \Bigg\{ - n \log c(\mu, P_{Z^+}) - \tag{28}$$

$$\frac{1}{2} \sum_{i=1}^{n} \phi\left(z^{(i)}, \mu, P_{Z^+}\right) \,:\, P_{22} \succ 0,\ P_{Z_1}^\star - P_{12} P_{22}^{-1} P_{12}^\top \succ 0 \Bigg\},$$

and

$$P_{Z^+}^\star = \begin{bmatrix} P_{Z_1}^\star & P_{12}^\star \\ P_{12}^{\star\top} & P_{22}^\star \end{bmatrix}, \text{ where } P_{Z^+} = \begin{bmatrix} P_{Z_1}^\star & P_{12} \\ P_{12}^\top & P_{22} \end{bmatrix}. \tag{29}$$

This is a non-convex program subject to positive definite constraints. The non-convexity is caused by $c(\mu, P_{Z^+})$. When $q = 1$, which corresponds to $Z_2(\delta)$ being a single monomial, the constraints in (28) are given by (27) with $P_{11} = P_{Z_1}^\star$. The algorithmic implementation of these ideas is presented next.

Mapping Reduction Algorithm for ML_δ*:* Assume that the data sequence $\mathcal{D}$ is available, and values for $d_{\max}$ and for the acceptance ratio $r > 1$ have been set. Furthermore, group the monomials of mapping $Z(\delta, d_{\max})$ to form $[Z_1, \ldots Z_k]$, and make $Z \leftarrow Z_1$.

```
$\langle \mu_Z^\star, P_Z^\star \rangle \leftarrow$ Equation (8).
**For** $i = 2, \ldots k$ **do**
    $J \leftarrow \mathcal{L}(\mu_Z^\star, P_Z^\star, \mathcal{D})$
    $Z^+ \leftarrow \{Z, Z_i\}$
    $\mu_{Z^+}^\star \leftarrow$ Equation (28) [5], $P_{Z^+}^\star \leftarrow$ Equation (29).
    **If** $\mathcal{L}(\mu_{Z^+}^\star, P_{Z^+}^\star, \mathcal{D}) > rJ$ **then**
        $\langle \mu_Z^\star, P_Z^\star \rangle \leftarrow \langle \mu_{Z^+}^\star, P_{Z^+}^\star \rangle$, $Z \leftarrow Z^+$
    **end if**
**end do**
```

[3] In contrast to (3), the one-argument notation $Z(\delta)$ will refer to an arbitrary polynomial mapping hereafter.

[4] Hereafter we will make Z a subindex of μ and P to highlight the dependency.

[5] Computing the matrix inverse using the developments in the next section yields computational savings.

The resulting parameters of the SN are $\mu = \mu^\star_Z$, and $P = P^\star_Z$, which correspond to the mapping $Z(\delta) = Z^+$. This formulation will be referred to as RML$_\delta$ hereafter. As expected, the performance of this SN is suboptimal, i.e., $\mathcal{L}(\mu^\star_{Z^+}, P^\star_{Z^+}, \mathcal{D}) \leq \mathcal{L}(\mu^\star, P^\star, \mathcal{D})$, where $\mu^\star$ and $P^\star$ are given by (8) for $Z = Z^+$.

5.2 Maximum Likelihood in Feature Space

The covariance of the data in feature space, $K > 0$, which is the inverse of $P^\star$ in (12), can be written as

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{12}^\top & K_{22} \end{bmatrix}, \tag{30}$$

where $K_{11} > 0 \in \mathbb{R}^{p\times p}$, $K_{12} \in \mathbb{R}^{p\times q}$ and $K_{22} \in \mathbb{R}^{q\times q}$. As before, we will consider the baseline mapping $Z_1(\delta)$, and the augmented mapping $Z^+(\delta) = [Z_1(\delta), Z_2(\delta)]$. The SN parameter $P^\star_{Z^+}$ can be derived from the corresponding K to obtain

$$P^\star_{Z^+} = \begin{bmatrix} P^\star_{Z_1}\left(I + K_{12}\lambda^{-1}K_{12}^\top P^\star_{Z_1}\right) & -P^\star_{Z_1}K_{12}\lambda^{-1} \\ -\lambda^{-1}K_{12}^\top P^\star_{Z_1} & \lambda^{-1} \end{bmatrix}, \tag{31}$$

where $P^\star_{Z_1} = K_{11}^{-1}$ and $\lambda = K_{22} - K_{12}^\top P^\star_{Z_1} K_{12}$ is the Schur complement of the block K_{11}. Equation (31) reduces the computational complexity of the SN estimation by leveraging $P^\star_{Z_1}$ to calculate $P^\star_{Z^+}$, i.e., by requiring the inversion of a $q \times q$ matrix instead of the inversion of a $p \times p$ matrix. Recall from Section 3 that this matrix inversion is the numerical bottle neck of ML$_z$ and ML$_z^+$. As before, the sequential implementation of (31) allows building a polynomial mapping that only contains the monomials that significantly increase (7).

Mapping Reduction Algorithm for ML$_z^+$: Starting[6] from the same setup of the previous algorithm proceeds as follows:

```
(1)  $\mu^\star_Z \leftarrow$ Equation (11), and $P^\star_Z \leftarrow$ Equation (12).
(2)  $\gamma^\star_{Z^+} \leftarrow$ Equation (13) given $\mu^\star_Z$ and $P^\star_Z$.
(3)  For $i = 2, \ldots k$ do
(4)      $J \leftarrow \mathcal{L}(\mu^\star_Z, \gamma^\star_{Z^+} P^\star_Z, \mathcal{D}),\ Z^+ \leftarrow \{Z, Z_i\}$.
(5)      $\mu^\star_{Z^+} \leftarrow$ Equation (11).
(6)      $P^\star_{Z^+} \leftarrow$ Equation (31) given $P^\star_Z$.
(7)      $\gamma^\star_{Z^+} \leftarrow$ Equation (13) given $\mu^\star_{Z^+}$ and $P^\star_{Z^+}$.
(8)      If $\mathcal{L}(\mu^\star_{Z^+}, \gamma^\star_{Z^+} P^\star_Z, \mathcal{D}) > rJ$ then
(9)          $\mu^\star_Z \leftarrow \mu^\star_{Z^+},\ P^\star_Z \leftarrow P^\star_{Z^+},\ Z \leftarrow Z^+$
(10)     end if
(11) end do
```

The parameters of the desired SN are $\mu = \mu^\star_{Z^+}$ and $P = \gamma^\star_{Z^+} P^\star_{Z^+}$, which correspond to the mapping $Z(\delta) = Z^+$. This formulation will be referred to as RML$_z^+$ hereafter.

Example 3: Next we apply the mapping reduction techniques to the data sequence in Example 1. We consider ML$_\delta$, ML$_z$ and ML$_z^+$, along with their reduced mapping variants, where the terms sequentially added to the map are single monomials in lexicographic order and $r = 1$. Figure 5 shows the likelihood of the data as a function of the dimension of the feature space n_z. Note that points of the reduced mapping variants on the same abscissa are not in the same feature space. ML$_\delta$ outperforms all other formulations by a significant margin for all n_z values whereas the performance of ML$_z$ degrades rapidly as n_z increases. Their reduced mapping variants, RML$_\delta$ and RML$_z$, improve monotonically until $n_z = 8$ and $n_z = 7$ respectively. Out of these two approaches however, only RML$_z$ improves its baseline formulation.

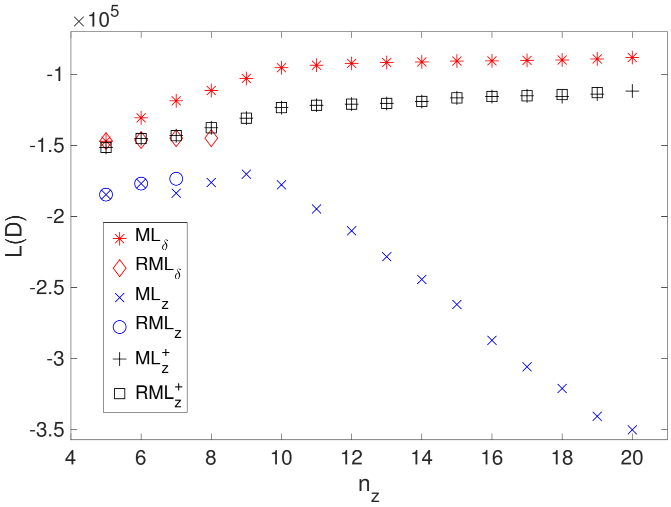

Fig. 5. Dimension of feature space vs. log-likelihood of the data.

The SNs resulting from the ML$_z^+$ are significantly better than those derived from ML$_z$. Furthermore, RML$_z^+$ yields slightly better results than ML$_z^+$ for $n_z = 18$ and $n_z = 19$. The SN did not improve beyond such a value. Note that the reduced SNs depend on the assumed grouping and ordering of $Z(\delta, d_{\max}) = [Z_1, \ldots Z_k]$, so other groups might lead to better SNs. Figure 5 also shows that the likelihood of the data corresponding to ML$_\delta$ and ML$_z^+$ starts plateauing at $n_z = 10$. As such, a convergence analysis should be used to select the lowest-complexity SN that still models the data sequence well. The good performance of RML$_z^+$ along with its lower computational cost make it well suited for high-dimensional datasets.

REFERENCES

Colbert, B., Crespo, L., and M., P. (2020). Improved uncertainty quantification for sliced normal distributions of increasing degree and dimension. *American Control Conference*.

Crespo, L., Colbert, B., and Kenny, S. (2019a). On the quantification of aleatory and epistemic uncertainty using sliced-normal distributions. *Systems and Control Letters*, 134.

Crespo, L.G., Giesy, D., Kenny, S., and Deride, J. (2019b). A scenario optimization approach to system identification with reliability guarantees. *American Control Conference*.

Czado, C. (2019). Analyzing dependent data with vine copulas. *Springer International Publishing*.

Haff, I., R., H., et al. (2013). Parameter estimation for pair-copula constructions. *Bernoulli*.

Kjersti, A., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, 44(2).

Kurowicka, D. and Cooke, R. (2006). Uncertainty analysis with high dimensional dependence modeling. *John Wiley & Sons*.

Nelsen, R. (2007). An introduction to copulas. *Springer Science & Business Media*.

[6] Remove Steps 2 and 7 and make $\gamma^\star_{Z^+} = 1$ in Steps 4 and 8 to obtain RML$_z$.