A Convex Optimization Approach to Improving Suboptimal Hyperparameters of Sliced Normal Distributions

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

Abstract— **Sliced Normal (SN) distributions are a generalization of Gaussian distributions where the quadratic argument of the exponential is replaced with a sum of squares polynomial. SNs may be used to represent the distribution of a diverse set of random variables including multi-modal, non-symmetric, and skewed distributions. Unfortunately, the likelihood function of a SN includes a normalization constant and the inclusion of this normalization constant makes the likelihood a non-convex function of the hyperparameters which define the SN. In previous work, suboptimal fitting of the hyperparameters was performed by transforming the given data into a higher dimensional monomial basis and selecting the optimal hyperparameters of a Gaussian fit in this space. However, this approach did not account for the effect of lifting on the normalization constant. Indeed, it was observed that as the number of monomials is increased the likelihood of the Sliced Normal can decrease. In this paper, we increase the likelihood of Sliced Normals found using the previous method by developing a convex formulation which scales the covariance matrix of the Gaussian fit such that the likelihood of the Sliced Normal is maximized. The result is significant improvements of the log likelihood of fitted SN distributions, including a significant increase, especially for problems with 500+ monomials.**

I. Introduction

The characterization of uncertainty in measured data is of significance for system identification, robust analysis [1], and robust controller synthesis [2]. Poor characterization of uncertainty can lead to either conservative or unreliable system analysis and controller design. For example, in [3], it was shown that for a particular stable uncertain system: when the uncertainties in system parameters were assumed to be independent, instability occurred in 20% of the simulations; whereas, when dependencies between variables were correctly modeled, all simulations were stable.

Methods of uncertainty characterization that account for parameter dependencies often assume Gaussian distributions:

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

However, Gaussian distributions are symmetric and unimodal - implying a linear dependence of the parameters δ. In [4] the alternative Sliced Normal (SN) class of distributions were proposed for modeling complex nonlinear parameter dependencies - (See Eqn. (2)). In this paper we propose an improved algorithm for the selection of the optimal hyperparameters (μ, P) for these SN distributions. Specifically, let $\delta\in\mathbb{R}^n$ be a random variable and $Z_d:\mathbb{R}^n\to\mathbb{R}^q$ denote the vector of monomials of degree less than d but greater than 0 of δ. Then the SN distribution is defined as

$$f(\delta;\mu,P)=\begin{cases}\frac{e^{-\frac{(Z_d(\delta)-\mu)^\top P(Z_d(\delta)-\mu)}{2}}}{c(\mu,P)} & \text{for } \delta\in\Delta\\ 0 & \text{else,}\end{cases} \tag{2}$$

where $\Delta\subset\mathbb{R}^n$ is the support set, c is the normalization constant, and $\mu\in\mathbb{R}^q$, and $P\in\mathbb{R}^{q\times q}$ are hyperparameters of the distribution.

The Non-convex SN Formulation:
Given a data sequence $\mathcal{D}=\{\delta^{(1)},\dots,\delta^{(m)}\}$ comprised of IID samples, we want to model the data with a SN by seeking a μ and P which maximize the likelihood of the data sequence $\mathcal{D}$. We may explicitly formulate this optimization problem as

$$\max_{\mu,P}\prod_{\delta\in\mathcal{D}}\frac{e^{-\frac{(Z_d(\delta)-\mu)^\top P(Z_d(\delta)-\mu)}{2}}}{c(\mu,P)}. \tag{3}$$

Unfortunately, however, this native formulation of the problem is non-convex and is thus difficult to solve when n, the number of parameters in δ, or q, the number of monomials in the monomial basis, is large, see [5]. In [4] we used a convex approximation of the numerical integration constant ($c(\mu,P)$), yielding a convex approximation of the original non-convex optimization problem (The Baseline SN Method). We found that when q (the number of monomials) is small this convex approximation method generated SNs that significantly outperformed Gaussian distributions. However, we also observed that for sufficiently large q, the performance of the convex approximation algorithm can actually decrease - implying the accuracy of the approximation decreases as the number of hyper-parameters increases.

In this paper, we show that for a given suboptimal μ and P obtained from the convex approximation algorithm, we can substantially increase the likelihood of the data by either scaling P or iteratively scaling subsets of P. Moreover, and significantly, we find that this search for the optimal scaling factors which maximizes the likelihood of the data sequence is convex.

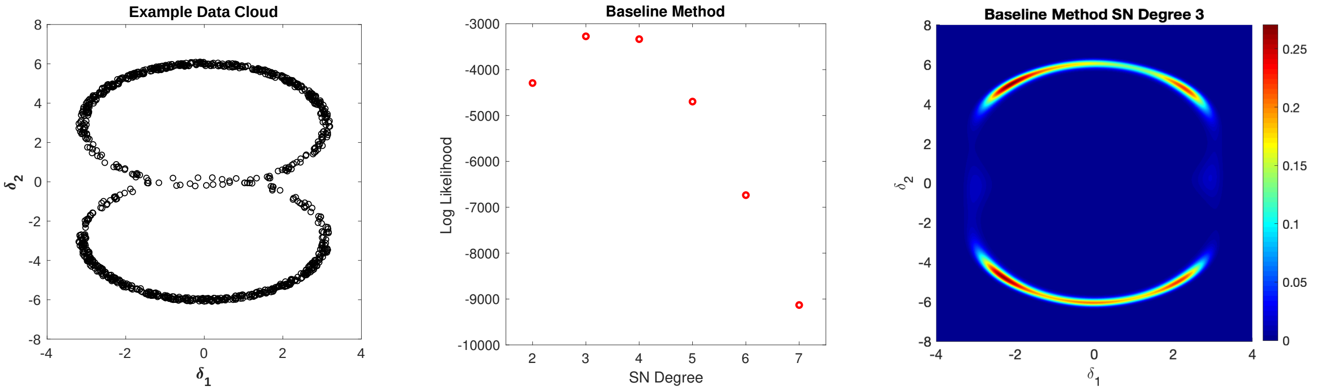
II. Notation

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

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

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

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



(a) Data cloud generated in polar coordinates by sampling 1000 points from a data generating mechanism[1].

(b) Likelihood of the SN on the data sequence $\mathcal{D}$ for varying degree using the baseline method.

(c) The SN with hyperparameters optimized using the baseline method that had the highest likelihood to the data given in subfigure (a).

Fig. 1: Subfigures a-c show the data sequence $\mathcal{D}$, the log likelihood of SN's found using the baseline method, solving Optimization Problem (4), for varying degree, and the joint PDF of the degree SN that had the maximal likelihood.

0, where $q = \binom{n+d}{n} - 1$. We denote the Hadamard product with $\circ$.

III. Optimized Sliced Normals Using an Approximate Normalization Constant

In this section, we give a summary of the previous baseline method for SN density estimation.

The Baseline SN Method:
This method uses a convex approximation of the integration constant in Optimization Problem (3). Specifically, we approximate $c(\mu, P)$ in Eq. (2) with the normalization constant of a multivariate gaussian where $c = (2\pi)^{q/2}\sqrt{|P^{-1}|}$. This approximation yields the following convex formulation:

$$\max_{P\in\mathbb{S}^+,\ \mu\in\mathbb{R}^q}\left\{\log\prod_{\delta\in\mathcal{D}}\frac{e^{-\frac{(Z_d(\delta)-\mu)^\top P(Z_d(\delta)-\mu)}{2}}}{(2\pi)^{q/2}\sqrt{|P^{-1}|}} : P\succ 0\right\}. \tag{4}$$

For a fixed degree d, Optimization Problem (4) admits an analytical solution where the optimal hyperparameters are,

$$\mu^* = \frac{1}{m}\sum_{i=1}^{m} Z_d(\delta^{(i)}), \quad P^* = \Sigma^{-1},$$

and $\Sigma = \frac{1}{m}\sum_{i=1}^{m}(Z_d(\delta^{(i)}) - \mu^*)(Z_d(\delta^{(i)}) - \mu^*)^\top$.

While the baseline method works well when the P matrix of the SN is small, we find that the likelihood of the fitted data sequence $\mathcal{D}$ can actually decrease when a large monomial basis is selected for the SN. For example, in Fig. 1(b) we compare SNs of degree d where the monomial basis consists of all monomials of degree d or less. One would expect the larger degree SNs to perform better because they have more hyperparameters to fit to the data. However, the SN with the best likelihood for the data sequence in Fig. 1(a) is of degree 3, with higher degree SN's having a smaller likelihood value. In Fig. 1(c) we see that baseline method allocates likelihood poorly in the areas with a low concentration of points, but does place high likelihood on areas with a high concentration of points.

In the following section we propose a new convex formulation of the problem which seeks to increase the likelihood of the data by scaling the hyperparameter P that results from the baseline method. This practice, in turn, has been observed to increase the covariance of the SN in physical space and generates SN's with higher likelihood than those generated by the baseline method.

IV. Optimized Sliced Normals By Scaling The P Hyperparameter

In this section, we propose a convex optimization problem that maximizes the likelihood of the SN for a given data sequence by scaling the suboptimal P matrix obtained from the baseline method without using a convex approximation for the function c.

In this case, we assume that a degree d has been selected and we are given previously selected values of the hyperparameters $P \in \mathbb{R}^{q\times q}$ and $\mu \in \mathbb{R}^q$, presumably found using Optimization Problem (4). We now consider SN 'candidates' of the following form,

$$f(\delta; \gamma P, \mu, \Delta) = \begin{cases} \frac{e^{-\phi(\delta,\mu,\gamma P)}}{c(\mu,\gamma P)} & \text{for } \delta \in \Delta \\ 0 & \text{else,} \end{cases} \tag{5}$$

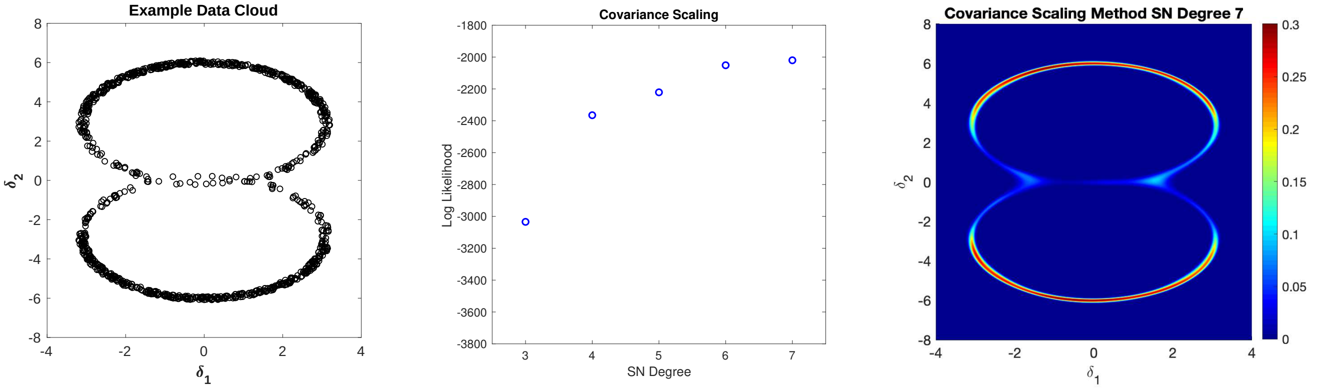
where P and μ are fixed, $\Delta \subset \mathbb{R}^n$ is the support set of the SN, and

$$\phi(\delta, \mu, P) = \frac{(Z_d(\delta) - \mu)^\top P(Z_d(\delta) - \mu)}{2}. \tag{6}$$

Our goal then, is to find a solution to the following optimization problem,

$$\max_{\gamma\in\mathbb{R}^+}\left\{\log\prod_{\delta\in\mathcal{D}}\frac{e^{-\phi(\delta,\mu,\gamma P)}}{c(\mu,\gamma P)}\right\}, \tag{7}$$

[1] Data set is generated by uniformly sampling between $[2.8, 3.2]$ to define the radius and using a normal distribution centered at 90 degrees with standard deviation of 1.3 to define the angle of the points in polar coordinates. Half of the points are translated 2.9 units in δ_2 while the other half are reflected over the δ_1 axis and then translated -2.9 units.



(a) Data cloud generated in polar coordinates by sampling 1000 points from a data generating mechanism.

(b) Likelihood of the SN on the data sequence $\mathcal{D}$ for varying degree using the covariance scaling method.

(c) The SN with hyperparameters optimized using the covariance scaling method that had the highest likelihood to the data given in subfigure (a) of the degrees tested.

Fig. 2: Subfigures a-c show the data sequence $\mathcal{D}$, the log likelihood of SN's optimized with the covariance scaling method for varying degree, and the joint PDF of the degree SN that had the maximal likelihood.

which optimizes the likelihood of the candidate SN using an accurate representation of the normalization constant c. Our approach to constructing an accurate convex representation of the normalization constant c are presented in Subsection IV-A.

A. *Numerical Calculation of the Normalization Constant*

To construct an accurate convex representation of c, we will use Monte Carlo integration techniques. First, we take a uniform sampling of the set Δ, as $s^{(i)}$ for $i = 1, ..., b$. If the volume of Δ is V, then we may now approximate the normalization constant of the SN as,

$$c_\Delta(\mu, \gamma P) = \int_\Delta e^{-\phi(\delta,\mu,\gamma P)} d\delta \approx \frac{V}{b}\sum_{i=1}^{b} e^{-\phi(s^{(i)},\mu,\gamma P)}. \tag{8}$$

Note that this approximation requires the volume of Δ, which is the support of the SN.

Note however that if Δ does not tightly enclose the data, when n is large the likelihood of most of the $s^{(i)}$ samples might be close to zero, and our approximation may be inaccurate. This deficiency can be mitigated by increasing b, sampling uniformly over a smaller data-containing set of known volume, or using an importance sampling technique [6].

In this paper we estimate c by sampling uniformly over an ellipsoidal set that tightly encloses the data. This set is calculated by using the worst-case likelihood formulation presented in [4]. The uniformly distributed $s^{(i)}$ samples are obtained by using the technique proposed in [8]. The volume of this set can be considerably smaller than the volume of a hyper-cube when the dimension, n, increases.

B. *Optimizing the Maximal Likelihood of a SN by Scaling the P Hyperparameter*

For $\delta \in \Delta$, our candidate function with the numerical approximation c_Δ of the integration constant, c, is,

$$f(\delta; \gamma P, \mu, \Delta) = \frac{e^{-\phi(\delta,\mu,\gamma P)}}{c_\Delta(\mu, \gamma P)}. \tag{9}$$

We may take the log of our objective function and find,

$$\begin{aligned}\log \prod_{\delta\in\mathcal{D}} \frac{e^{-\phi(\delta,\mu,\gamma P)}}{c_\Delta(\mu,\gamma P)} &= \sum_{\delta\in\mathcal{D}} \log \frac{e^{-\gamma\phi(\delta,\mu,P)}}{c_\Delta(\mu,\gamma P)},\\ &= m\log\Bigg(\frac{1}{c_\Delta(\mu,\gamma P)}\Bigg) - \gamma\sum_{\delta\in\mathcal{D}}\phi(\delta,\mu,P)\end{aligned} \tag{10}$$

where, for a given sampling $s^{(i)}$ of Δ,

$$c_\Delta(\mu, P) = \frac{V}{b}\sum_{i=1}^{b} e^{-\gamma\phi(s^{(i)},\mu,P)}, \tag{11}$$

where V is the volume of Δ.

The Covariance Scaling SN Method:

Inserting Eq. (11) into Eq. (10) leads to,

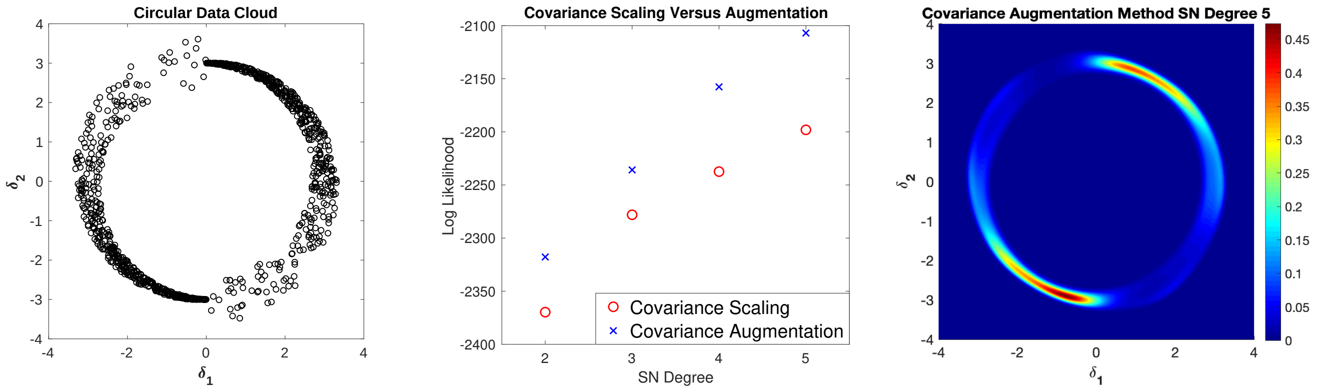
$$\begin{aligned}&m\log\left(\frac{1}{c_\Delta(\mu,\gamma P)}\right) - \gamma\sum_{\delta\in\mathcal{D}}\phi(\delta,\mu,P)\\ &= m\log\left(\frac{b}{V\sum_{i=1}^{b} e^{-\gamma\phi(s^{(i)},\mu,P)}}\right) - \gamma\sum_{\delta\in\mathcal{D}}\phi(\delta,\mu,P)\\ &= m\log(\frac{b}{V}) - m\log\Bigg(\sum_{i=1}^{b} e^{-\gamma\phi(s^{(i)},\mu,P)}\Bigg) - \gamma\sum_{\delta\in\mathcal{D}}\phi(\delta,\mu,P).\end{aligned}$$

Since the first term is a constant, we may rewrite Optimization Problem (7) as,

$$\max_{\gamma\in\mathbb{R}^+}\left\{-m\log\left(\sum_{i=1}^{b} e^{-\gamma\phi(\delta,\mu,P)}\right) - \gamma\sum_{\delta\in\mathcal{D}}\phi(\delta,\mu,P)\right\}. \tag{12}$$

Optimization Problem (12) is now convex, a fact we prove in the following section.

Since the optimization problem is convex, any local maximum is a global maximum and we may therefore use a gradient based algorithm to seek the global optimum. The need for a gradient-based search is the result of having to estimate the integration constant numerically.



(a) Data cloud generated in polar coordinates by sampling 1000 points from a data generating mechanism[2].

(b) Likelihood of the SN on the data sequence $\mathcal{D}$ for varying degree using the covariance scaling versus the covariance augmentation method.

(c) The SN with hyperparameters optimized using the covariance augmentation method that had the highest likelihood to the data given in subfigure (a).

Fig. 3: Subfigures a-c show the data sequence $\mathcal{D}$, the log likelihood of SN's optimized with the covariance scaling and augmentation method for varying degree, and the joint PDF of the covariance augmented degree five SN that had the maximal likelihood.

In Fig. 2(b) we see that the SN with the best likelihood for the data sequence in Fig. 2(a) is of degree 7 and that, unlike the baseline method, increasing the degree increases the likelihood of the SN. Fig. 2(c) shows that the SN of degree 7 does model the dataset better than the SN resulting from the baseline approach.

In this section we have assumed that γ is a scalar factor that affects the hyperparameter P. In the following section we will now perform a multivariate scaling of P.

V. Optimized Sliced Normals by Iterative Scaling of Subsets of the P Hyperparameter

In this section we develop a multivariate extension of the covariance scaling approach described earlier, by iteratively modifying the P matrix from the Baseline method. Again we assume that a degree d has been selected and we have a $P \in \mathbb{R}^{q\times q}$, and a $\mu \in \mathbb{R}^q$, that define the hyperparameters of a SN found using the baseline method. Then we consider SN 'candidates' of the following form,

$$f(\delta;\mu,\Gamma\circ P) = \begin{cases} \frac{e^{-\phi(\delta,\mu,\Gamma\circ P)}}{c(\mu,\Gamma\circ P)} & \text{for } \delta\in\Delta \\ 0 & \text{else,} \end{cases} \tag{13}$$

where $\circ$ is the Hadamard product and the hyperparameter is $\Gamma \in \mathbb{R}^{q\times q}$ (instead of the scalar γ in the previous section). Note that we do not constrain Γ to be positive definite, differing from the original construction of the Sliced Normal distribution which required the polynomial to be sum-of-squares. However, $f(\delta;\mu,\Gamma\circ P)$ is still globally positive and the Sliced Normal is thus still a valid distribution.

Our goal then, is to find a solution to the following optimization problem,

$$\max_{\Gamma\in\mathbb{R}^{q\times q}} \left\{ \log \prod_{\delta\in\mathcal{D}} \frac{e^{-\phi(\delta,\mu,\Gamma\circ P)}}{c(\mu,\Gamma\circ P)} \right\}, \tag{14}$$

where we may again use the accurate convex approximation c_Δ to optimize the likelihood. As in the previous method we have,

$$\begin{aligned} \log \prod_{\delta\in\mathcal{D}} \frac{e^{-\phi(\delta,\mu,\Gamma\circ P)}}{c(\mu,\Gamma\circ P)} &= \sum_{\delta\in\mathcal{D}} \log \frac{e^{-\phi(\delta,\mu,\Gamma\circ P)}}{c(\mu,\Gamma\circ P)}, \\ &= m\log\!\left(\frac{1}{c(\mu,Q)}\right) - \sum_{\delta\in\mathcal{D}} \phi(\delta,\mu,Q). \end{aligned} \tag{15}$$

where $Q = \Gamma\circ P$.

Given V, the volume of a Δ that has been uniformly sampled, we may again approximate c as,

$$c_\Delta(\mu,Q) = \frac{V}{b}\sum_{i=1}^{b} e^{-\phi(s^{(i)},\mu,Q)},$$

where V is the volume of Δ. We may then further simplify Eq. (15) to,

$$\begin{aligned} & m\log\!\left(\frac{1}{c_\Delta(\mu,Q)}\right) - \sum_{\delta\in\mathcal{D}} \phi(\delta,\mu,Q) \\ &\quad = m\log\left(\frac{k}{V\sum_{i=1}^{k} e^{-\phi(s^{(i)},\mu,Q)}}\right) - \sum_{\delta\in\mathcal{D}} \phi(\delta,\mu,Q) \\ &\quad = c_V - m\log\!\left(\sum_{i=1}^{k} e^{-\phi(s^{(i)},\mu,Q)}\right) - \sum_{\delta\in\mathcal{D}} \phi(\delta,\mu,Q), \end{aligned}$$

where $c_V = m\log(\frac{k}{V})$. Since c_V is a constant, we may rewrite Optimization Problem (14) as,

$$\max_{\Gamma\in\mathbb{R}^{q\times q}} \left\{ -m\log\left(\sum_{\delta\in\mathcal{S}} e^{-\phi(\delta,\mu,\Gamma\circ P)}\right) - \sum_{\delta\in\mathcal{D}} \phi(\delta,\mu,\Gamma\circ P) \right\}. \tag{16}$$

[2] Data set is generated by sampling 500 points a normal distribution centered at $\frac{\pi}{2}$ radians with standard deviation of 1.3 to define the angle of the points in polar coordinates, the radius of the points is 3 plus a value randomly sampled between 0 and .2 multiplied by the angle of the point minus $\frac{\pi}{2}$. Any point with a δ_1 value less than 0 are reflected over the δ_1 axis. This data set is sampled again, but these points are additionally reflected over the δ_1 and δ_2 axes.

Next we will prove convexity of a slightly more general class of optimization problem. This result will then be used to show that Optimization Problem (12) is convex and that Optimization Problem (16) is multi-convex.

Theorem 1: For any $r \in \mathbb{N}$, $m \in \mathbb{N}$, $\alpha \in \mathbb{R}^m$, $\beta \in \mathbb{R}^m$, $\kappa \in \mathbb{R}^n$, and $\psi \in \mathbb{R}^n$, the function,

$$f(\gamma) = -m \log\left(\sum_{i=1}^{r} e^{-\gamma\alpha_i + \beta_i}\right) - \sum_{j=1}^{m} \gamma\kappa_j + \psi_j$$

is concave with respect to $\gamma \in \mathbb{R}$.

Proof:

We will prove that the second derivative of $f(\gamma)$ is negative for all values of γ. The first derivative of $f(\gamma)$ is,

$$f' = -m \frac{\sum_{i=1}^{r} -\alpha_i e^{-\gamma\alpha_i+\beta_i}}{\sum_{i=1}^{r} e^{-\gamma\alpha_i+\beta_i}} - \sum_{j=1}^{m} \kappa_j,$$

and the second derivative f'' is,

$$\frac{n(\gamma)}{d(\gamma)} = \frac{-m\left(\sum\limits_{i=1}^{r} \alpha_i^2 a_i(\gamma)\right)\left(\sum\limits_{i=1}^{r} a_i(\gamma)\right) - \left(\sum\limits_{i=1}^{r} \alpha_i a_i(\gamma)\right)^2}{\left(\sum\limits_{i=1}^{r} a_i(\gamma)\right)^2},$$

where

$$a_i(\gamma) = e^{-\gamma\alpha_i+\beta_i} > 0 \quad \forall\ \gamma.$$

Clearly $d(\gamma)$ is positive so we must show that $n(\gamma)$ is negative to prove concavity of $f(\gamma)$. We have that,

$$\left(\sum_{i=1}^{r}\alpha_i^2 a_i(\gamma)\right)\left(\sum_{i=1}^{r} a_i(\gamma)\right) = \sum_{i=1}^{r} \alpha_i^2 a_i(\gamma)^2 + \sum_{i=1}^{r}\sum_{j=i+1}^{r} (\alpha_i^2 + \alpha_j^2) a_i(\gamma) a_j(\gamma),$$

and

$$\left(\sum_{i=1}^{r} \alpha_i a_i(\gamma)\right)^2 = \sum_{i=1}^{r} \alpha_i^2 a_i(\gamma)^2 + \sum_{i=1}^{r}\sum_{j=i+1}^{r} 2\alpha_i\alpha_j a_i(\gamma) a_j(\gamma),$$

therefore,

$$\begin{aligned} n(\gamma) &= -m \sum_{i=1}^{r}\sum_{j=i+1}^{r} (\alpha_i^2 + \alpha_j^2) a_i(\gamma) a_j(\gamma) - 2\alpha_i\alpha_j a_i(\gamma) a_j(\gamma) \\ &= -m \sum_{i=1}^{r}\sum_{j=i+1}^{r} a_i(\gamma) a_j(\gamma)(\alpha_i^2 - 2\alpha_i\alpha_j + \alpha_j^2) \\ &= -m \sum_{i=1}^{r}\sum_{j=i+1}^{r} a_i(\gamma) a_j(\gamma)(\alpha_i - \alpha_j)^2, \\ &\leq 0. \end{aligned}$$

Since $n(\gamma)$ is always negative and $d(\gamma)$ is always positive, $f''(\gamma)$ is negative for all γ and $f(\gamma)$ is concave. ■

To prove that the objective function of Optimization Problem (12) is concave let,

$$\alpha_i = \phi(s^{(i)}, \mu, P), \quad \kappa_j = \phi(\delta^{(j)}, \mu, P),$$

and β_i, ψ_j be zero for all values of i and j.

To prove that Optimization Problem (16) is multi-convex, let $\nu \in \mathbb{N}^{2\times r}$ be a list of r elements of Γ that we wish to optimize. Then let,

$$M_{i,j}^{(\nu)} = \begin{cases} P_{i,j}\Gamma_{i,j} & (i,j) \in \nu \\ 0 & \text{else} \end{cases}, \quad N^{(\nu)} = \Gamma \circ P - M^{(\nu)}$$

and

$$\alpha_i = \phi(s^{(i)}, \mu, M^{(\nu)}), \quad \beta_i = \phi(s^{(i)}, \mu, N^{(\nu)}) \text{ for all i and,}$$

$$\kappa_j = \phi(\delta^{(j)}, \mu, M^{(\nu)}), \quad \psi_j = \phi(\delta^{(j)}, \mu, N^{(\nu)}) \text{ for all j.}$$

Then, by Theorem 1, the objective function is concave.

The Covariance Augmentation SN Method: Optimization Problem (16) is multi-convex meaning we may iteratively optimize Γ using convex optimization problems to approximate the optimal solution. First we select a basis $v^{(i)} \in \mathbb{R}^{q\times q} \quad \forall i = 1, ..., k$ for $\mathbb{R}^{q\times q}$. Then we may use Algorithm 1 to find Γ.

Algorithm 1 The Covariance Augmentation Method

```
Given: $v$, $P$, $\mu$, $\mathcal{D}$, $s$, $N$
$\Gamma = \mathbf{1}$, $i = 0$
while $i < N$ do
  for $i = 1, ..., k$ do
    Calculate $\alpha$, $\beta$, $\kappa$ and $\psi$ for the basis $v^{(i)}$
    $\gamma = \underset{\gamma}{\text{argmax}} -m\log\left(\sum_{i=1}^{r} e^{-\gamma\alpha_i+\beta_i}\right) - \sum_{j=1}^{m} \gamma\kappa_j + \psi_j$
    $\Gamma = \Gamma + \gamma v^{(i)}$
  end for
  i = i+1
end while
return $\Gamma$
```

In Fig. 3, we show a comparison between the covariance scaling method, and the covariance augmentation method. For this example we choose the canonical basis for $\mathbb{R}^{q\times q}$ - optimizing a single element of Γ at each iteration until every element has been updated five times (1050 optimization problems).

In numerical tests we will compare using the canonical basis to the following basis,

$$v_{k,j}^{(i)} = \begin{cases} 1 & \text{if } \min(k,j) = i \\ 0 & \text{else.} \end{cases} \tag{17}$$

The benefit of the basis in Eq. (17) is that the number of matrices in the basis increases linearly with the size of the monomial basis. The number of matrices in the canonical basis however increases as a quadratic with respect to the size of the monomial basis.

In Section VI we will compare the performance of the baseline, covariance scaling and covariance augmentation methods.

TABLE I: Comparison between the Baseline Method (B), the Covariance Scaling Method (C_S), the Covariance Augmentation Method using the basis in Eq. (17) (C_{A1}), and the Covariance Augmentation Method using the canonical basis (C_{A2})

Data Set	n	Degree	q	Log Likelihood (1×10^3)				Time (seconds)			
				B	C_S	C_{A1}	C_{A2}	B	C_S	C_A	C_{A2}
Eight (Fig. 1(a))	2	5	20	-4.6965	-2.2212	-2.1897	-2.1740	1.48	1.56	44.50	350.05
Eight (Fig. 1(a))	2	7	35	-9.1309	-2.0201	-2.0033	-1.9925	1.67	2.30	85.57	1939.80
Circular (Fig. 3(a))	2	5	20	-6.3547	-2.1981	-2.1284	-2.1069	0.86	1.53	46.85	353.21
Iris [7]	4	3	34	-1.7208	-0.2316	-0.2043	-0.2009	2.15	2.31	87.48	1754.80
Iris [7]	4	4	69	-3.4371	-0.0977	N/A	N/A	5.62	5.99	N/A	N/A
Seeds [7]	7	3	119	-6.0016	1.8081	N/A	N/A	16.9312	16.9757	N/A	N/A
Abalone [7]	8	4	496	-196.78	14.450	N/A	N/A	180.15	180.98	N/A	N/A

VI. Comparison Study

Here we use datasets from the UCI machine learning repository [7] and datasets shown in Fig. 1(a) and Fig. 3(a) to compare the performance of the three different methods with respect to the log likelihood of the data sequence. The results can be seen in Table I, where the log likelihood of the optimized SN on the data sequence are presented as well as the time taken to estimate the SN and calculate the normalization constant.

In the table we compare the Baseline Method, the Covariance Scaling Method, the Covariance Augmentation Method with the basis in Eq. (17) iterated 10 times, and the Covariance Augmentation Method with the canonical basis iterated 5 times.

The Δ selected for the datasets in Fig. 1(a) and Fig. 3(a) were hand selected and are equal to the range of axis values seen in the plots. The samples for the datasets taken from the UCI database [7], were generated using the method in [8], where the ellipsoidal set was a level set of the optimal worst case degree 1 SN using the optimization method in [4].

Note that the difference in time taken between the baseline method and the covariance scaling method is negligible. This is because the most computationally expensive part of the method is calculating the normalization constant, which is necessary for the baseline method as well. Therefore, there seem to be few, if any, circumstances where the baseline method would be preferred to the covariance scaling method. Note that all of these methods work for problems where the P matrix has greater than 1000 elements, and the covariance scaling method works for problems where the P matrix has over 240,000 elements.

Note that as q (the size of the monomial basis) increases, the difference between the log likelihood of the baseline method and the covariance scaling method also increases. In addition, as the size of the monomial basis increases the covariance augmentation methods also begin to have numerical difficulties and fail or take a significant amount of computational time. Thus the more computationally expensive covariance augmentation methods should be reserved for problems with a smaller monomial basis than the covariance scaling method.

VII. Conclusion

This paper proposes a convex optimization problem for improving sub-optimal Sliced Normal hyperparameters obtained by the algorithm proposed in [4]. The proposed algorithm takes the solution of the covariance matrix from the Baseline Method for a Sliced Normal from Eq. (4) and finds an optimal scaling factor γ which scales the entire P matrix, or find an optimal matrix Γ through iteratively optimizing subsets of the P matrix.

The single scaling method offers large improvements over the previous baseline method with negligible additional computational expense. In cases where computation time is unimportant, the more complex iterative method may be used which offers additional increases in log likelihood values when compared to the covariance scaling method.

The developments herein allow for the efficient characterization of the dependencies in datasets of larger dimension. Properly characterizing the dependencies of the data is instrumental in system identification, robust analysis, and robust controller synthesis.

References

[1] 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," in *42nd IEEE International Conference on Decision and Control*, Maui, HI, USA, Dec. 2003, pp. 3766–3771.

[2] G. Chesi, "LMI techniques for optimization over polynomials in control: a survey," *IEEE Transactions on Automatic Control*, vol. 55, pp. 2500–2510, Nov. 2010.

[3] L. G. Crespo, D. Giesy, S. Kenny, and J. Deride, "A scenario optimization approach to system identification with reliability guarantees," in *2019 American Control Conference (ACC)*, Philadelphia, PA, USA, Jul. 2019, pp. 2100–2106.

[4] B. K. Colbert, L. G. Crespo, and M. M. Peet, "A sum of squares optimization approach to uncertainty quantification," in *2019 American Control Conference (ACC)*, Philadelphia, PA, USA, Jul. 2019, pp. 5378–5384.

[5] L. G. Crespo, B. K. Colbert, S. P. Kenny, and D. P. Giesy, "On the quantification of aleatory and epistemic uncertainty using sliced-normal distributions," *Systems & Control Letters*, vol. 134, Oct. 2019.

[6] P. W. Glynn and D. L. Iglehart, "Importance sampling for stochastic simulations," *Management science*, vol. 35, pp. 1367–1392, Nov. 1989.

[7] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[8] J. Dezert and C. Musso, "An efficient method for generating points uniformly distributed in hyperellipsoids," in *Proceedings of the Workshop on Estimation, Tracking and Fusion: A Tribute to Yaakov Bar-Shalom*, Monterey, CA, USA, May 2001.