

A Convex Optimization Approach to Improving Suboptimal Hyperparameters of Sliced Normals

Brendon K. Colbert
Luis G. Crespo
Matthew M. Peet

NASA Langley Research Center
Arizona State University

July 3, 2020



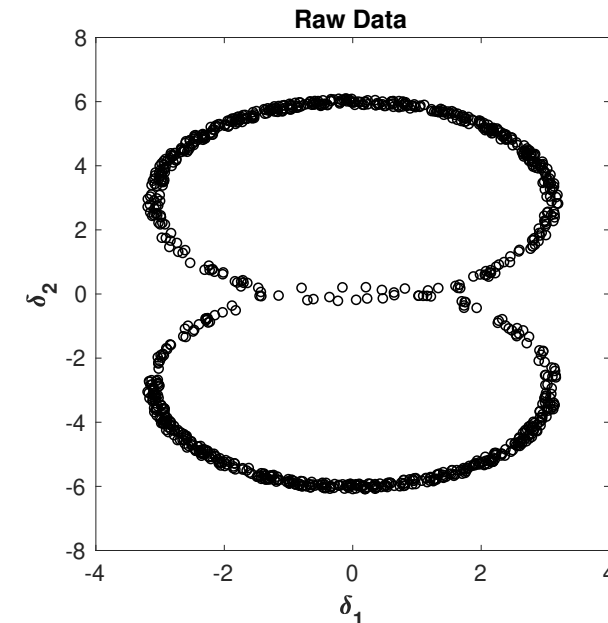
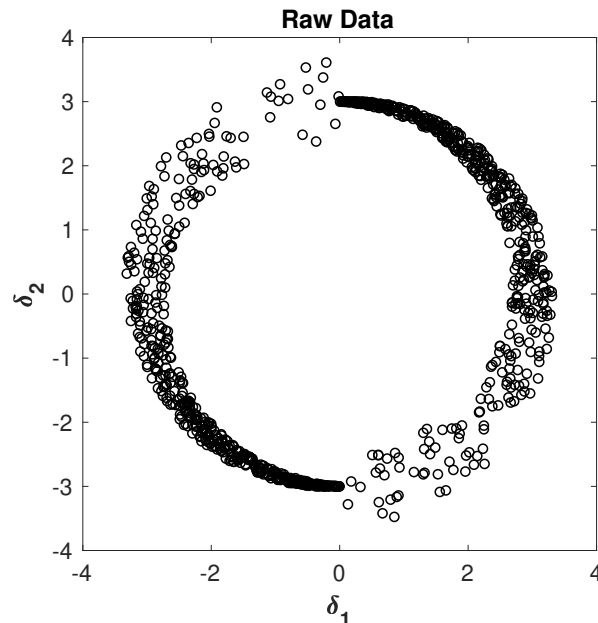
Characterizing Complex Data Generating Mechanisms

Given: Points $\mathcal{D} = \{\delta^{(i)}\}_{i=1}^m$

- Assume the points $\delta^{(i)}$ were drawn from a Data Generating Mechanism (DGM).
- The DGM has an underlying Probability Density Function (PDF) for all $\delta \in \Delta$.
- Each $\delta^{(i)}$ could represent the coefficients of an uncertain plant model.

We can model the data by:

- Directly modeling the PDF of the DGM.
- Finding a polynomial inequality that encloses the data (semi-algebraic sets).



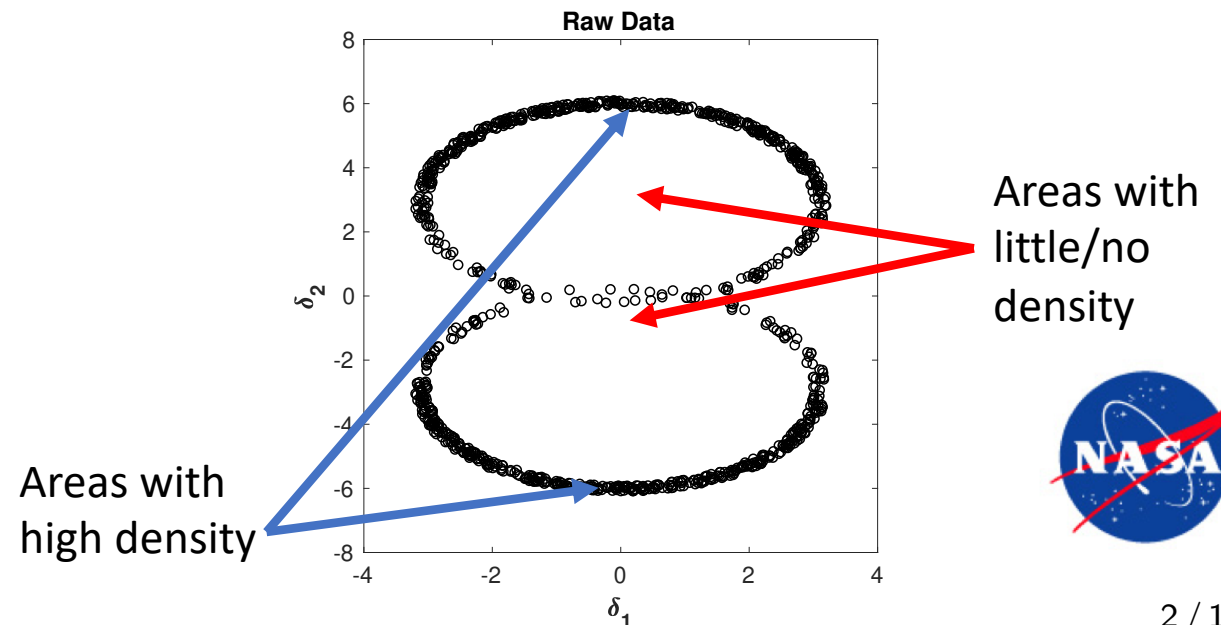
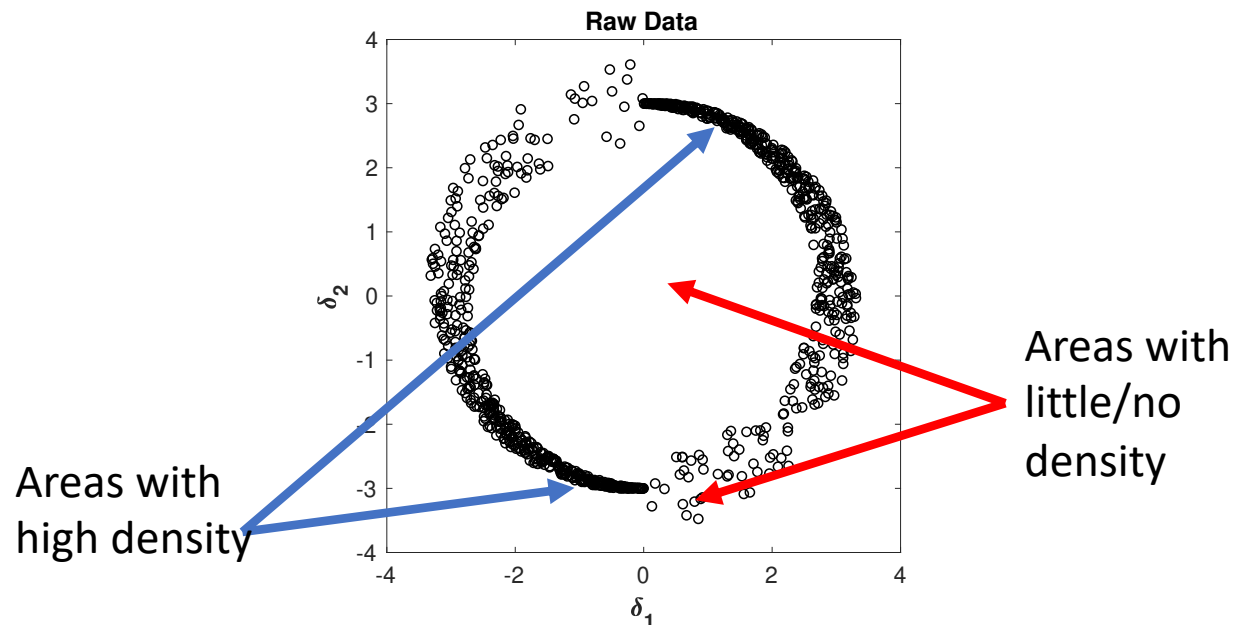
Characterizing Complex Data Generating Mechanisms

Given: Points $\mathcal{D} = \{\delta^{(i)}\}_{i=1}^m$

- Assume the points $\delta^{(i)}$ were drawn from a Data Generating Mechanism (DGM).
- The DGM has an underlying Probability Density Function (PDF) for all $\delta \in \Delta$.
- Each $\delta^{(i)}$ could represent the coefficients of an uncertain plant model.

We can model the data by:

- Directly modeling the PDF of the DGM.
- Finding a polynomial inequality that encloses the data (semi-algebraic sets).



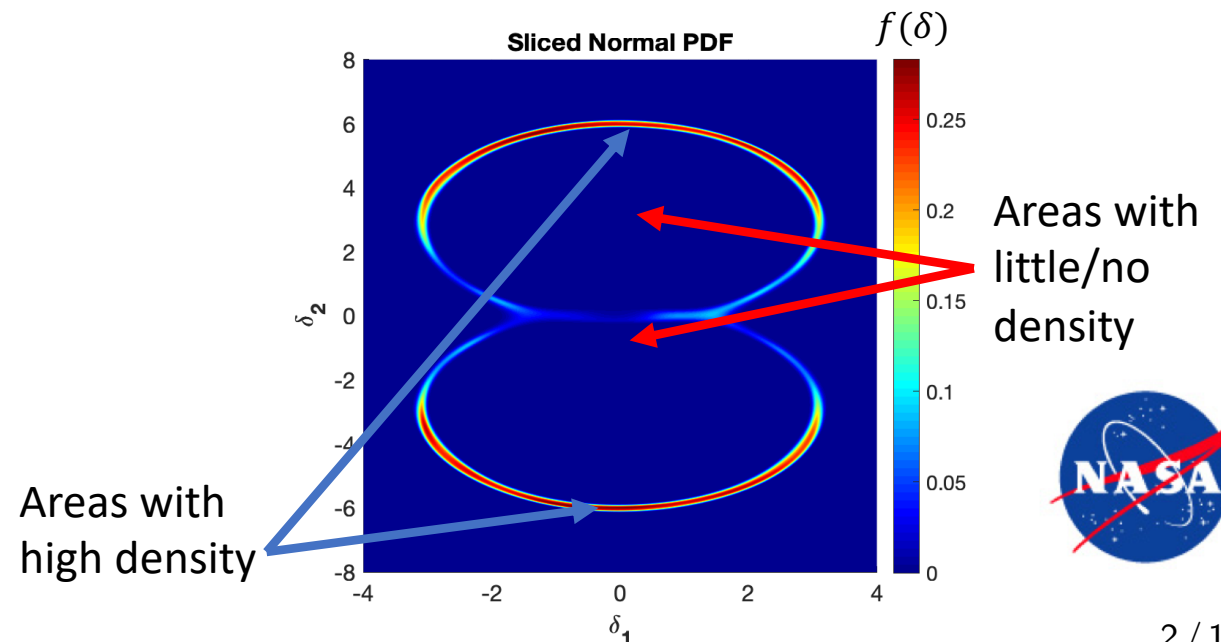
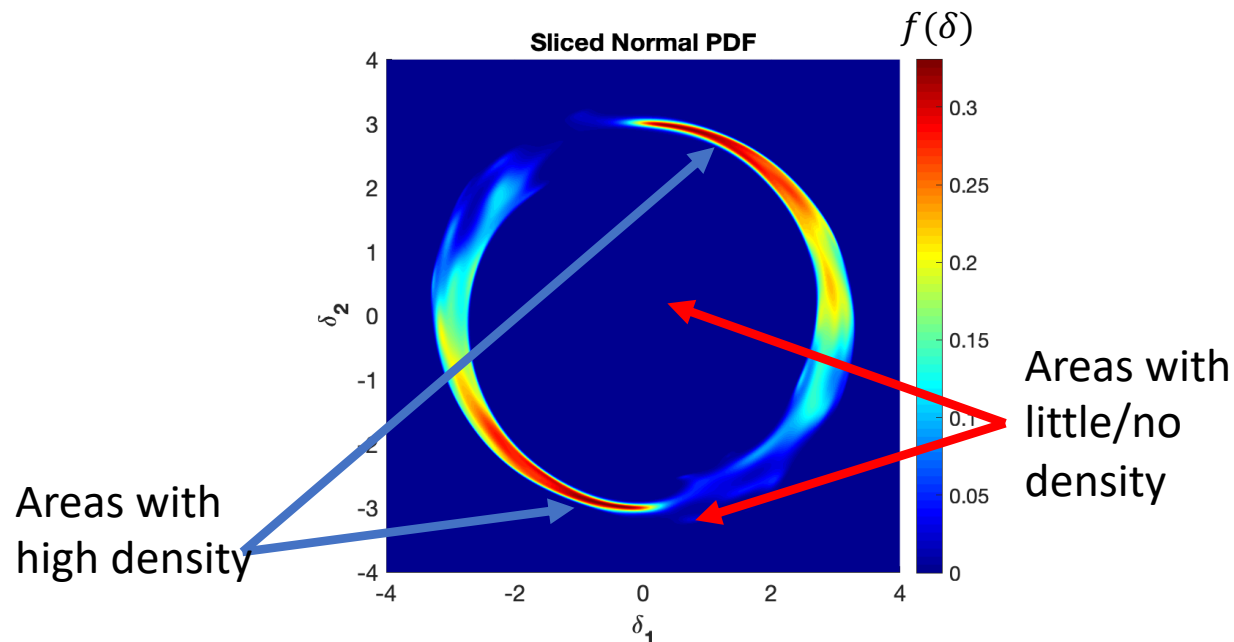
Characterizing Complex Data Generating Mechanisms

Given: Points $\mathcal{D} = \{\delta^{(i)}\}_{i=1}^m$

- Assume the points $\delta^{(i)}$ were drawn from a Data Generating Mechanism (DGM).
- The DGM has an underlying Probability Density Function (PDF) for all $\delta \in \Delta$.
- Each $\delta^{(i)}$ could represent the coefficients of an uncertain plant model.

We can model the data by:

- Directly modeling the PDF of the DGM.
- Finding a polynomial inequality that encloses the data (semi-algebraic sets).



Definition of our Sliced Normal PDF

Sliced Normal Probability Density Function

We say that f is a Sliced Normal (SN) Probability Density Function (PDF) if,

$$f(\delta; \underbrace{\mu, P, d}_{\text{Hyperparameters}}) = \frac{e^{-\phi(\delta; \mu, P, d)}}{\int_{\Delta} e^{-\phi(x; \mu, P, d)} dx}, \quad (1)$$

where $\phi(\delta; \mu, P, d) = \frac{1}{2}(Z_d(\delta) - \mu)^T P (Z_d(\delta) - \mu)$ is SOS for $P \succeq 0$ and $Z_d(\delta)$ are the monomials of degree d or less of δ .

For Example: $Z_2 \left(\begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} \right) = [\delta_1 \quad \delta_2 \quad \delta_1^2 \quad \delta_1\delta_2 \quad \delta_2^2]^T$

- We optimize the hyperparameters μ , and P (aka decision variables).
- $f(\delta; \mu, P, 1)$ is a Normal PDF.
- For $d > 1$ SN PDFs can characterize complex multi-modal distributions.

Question: How does one optimize μ and P given $\mathcal{D}$ and d ?



How to Select Optimal Hyperparameters

The Likelihood of the data $\mathcal{D}$

For a given PDF, $f(\delta)$, the likelihood of the data $\mathcal{D}$ is defined as

$$L_f(\mathcal{D}) = \prod_{i=1}^m f(\delta^{(i)}). \quad (2)$$

$L_f(\mathcal{D}) \geq L_g(\mathcal{D})$ implies that f is more likely to have generated the data, $\mathcal{D}$ than g .

- For a given $\mathcal{D}$ and degree d we want to solve,

$$\mu^*, P^* := \operatorname{argmax}_{\mu \in \mathbb{R}^q, P \in \mathbb{R}^{q \times q}} L_{f(\cdot, \mu, P, d)}(\mathcal{D})$$

- For a Normal PDF $d = 1$ and,

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

- P^* can be calculated in $O(n^3)$ where n is the dimension of δ .



Prior Work: Optimal Normal PDF in Lifted Space

Optimizing the Normal PDF in a Lifted Space

Recall $\mathcal{D} := \{\delta^{(i)}\}_{i=1}^m$ and let $\mathcal{D}_z := \{Z_d(\delta^{(i)})\}_{i=1}^m$, then we have that,

$$\max_{\substack{\mu \in \mathbb{R}^q \\ P \in \mathbb{R}^{q \times q}}} \overbrace{L_f(\cdot; \mu, P, d)(\mathcal{D})}^{\text{Likelihood}} = \underbrace{\max_{\substack{\mu \in \mathbb{R}^q \\ P \in \mathbb{R}^{q \times q}}} \frac{\overbrace{\int_{\mathbb{R}^q} e^{-\phi(z; \mu, P, 1)} dz}^{\text{Ratio of Normalization Constants}}}{\int_{\Delta} e^{-\phi(x; \mu, P, d)} dx} \underbrace{L_f(\cdot; \mu, P, 1)(\mathcal{D}_z)}_{\substack{\text{"Lifted" Likelihood} \\ \text{Convex}}}}_{\text{Non-Convex}}$$

We optimize the convex portion of $L_f(\cdot; \mu, P, d)(\mathcal{D})$, which is a Normal PDF in a “lifted space” to approximate the true solution.

- The solution to, $\max_{\substack{\mu \in \mathbb{R}^q \\ P \in \mathbb{R}^{q \times q}}} L_f(\cdot; \mu, P, 1)(\mathcal{D}_z)$, is

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

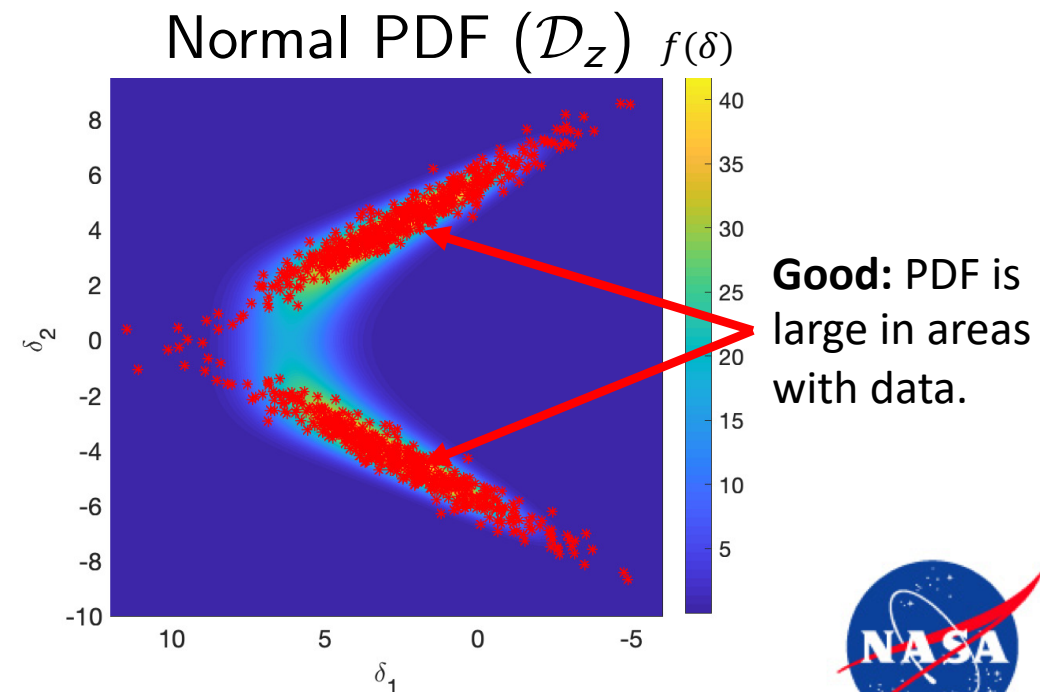
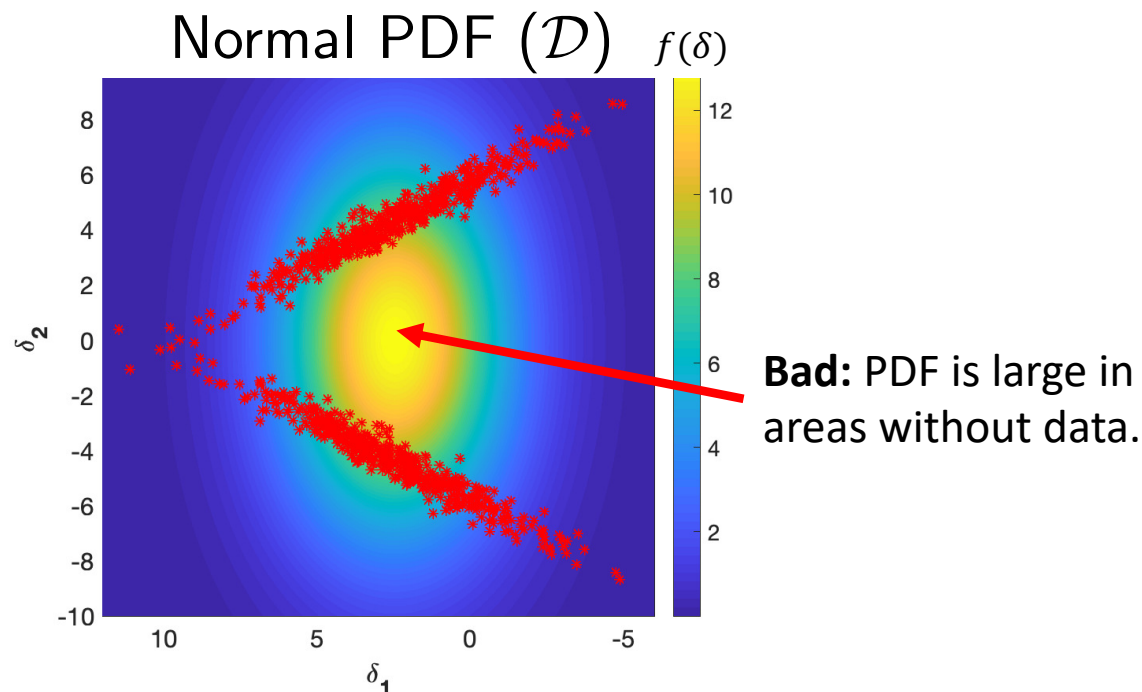


Prior Work: Optimal Normal PDF in Lifted Space

Optimizing the Normal PDF in a Lifted Space is Suboptimal (but works well)

The hyperparameters μ^* , P^* that optimize $L_{f(\cdot; \mu, P, 1)}(\mathcal{D}_Z)$ are suboptimal,

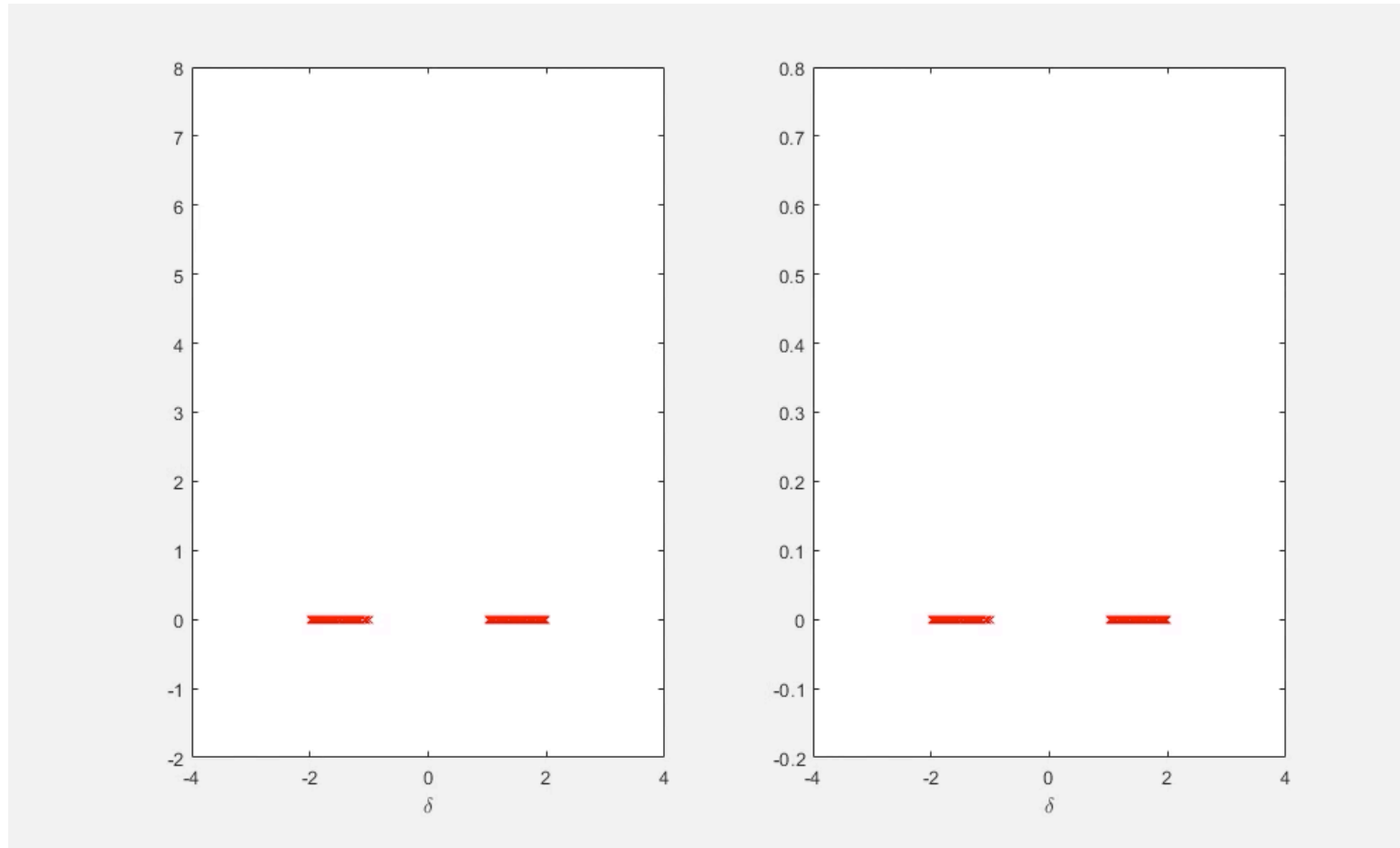
$$\max_{\substack{\mu \in \mathbb{R}^q \\ P \in \mathbb{R}^{q \times q}}} L_{f(\cdot; \mu, P, d)}(\mathcal{D}) \geq \frac{\int_{\mathbb{R}^q} e^{-\phi(x; \mu^*, P^*, 1)} dx}{\int_{\Delta} e^{-\phi(x; \mu^*, P^*, d)} dx} L_{f(\cdot; \mu^*, P^*, 1)}(\mathcal{D}_Z),$$



- The likelihood of the data is lower for the Normal PDF on the left than the suboptimal SN PDF on the right.



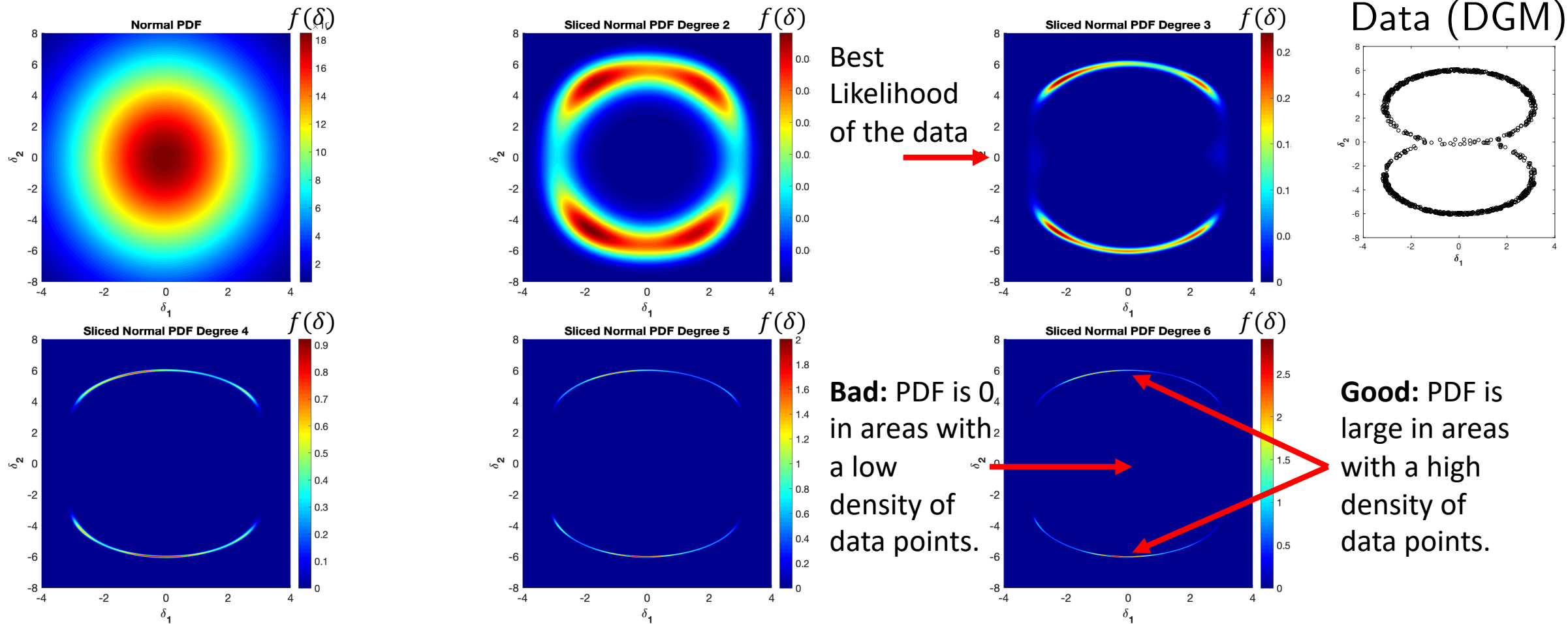
Prior Work: Visualizing the Lifted Space



- 1 Lift the data to a higher dimensional space e.g. $\delta^{(i)} \rightarrow z^{(i)} = \begin{bmatrix} \delta^{(i)} \\ (\delta^{(i)})^2 \end{bmatrix}$.
- 2 Find the μ^* and P^* of the optimal Normal PDF, $L_{f(\cdot; \mu, P, 1)}(\mathcal{D}_z)$.
- 3 Use μ^* and P^* for the SN PDF $f(\delta; \mu^*, P^*, d)$.



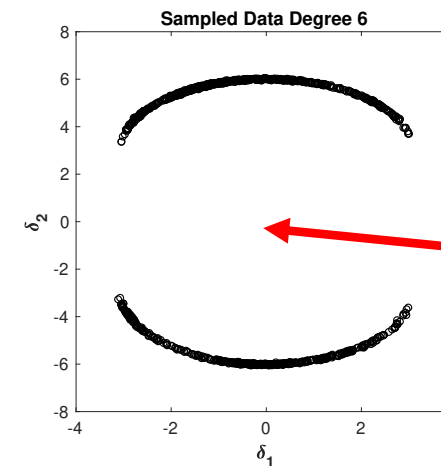
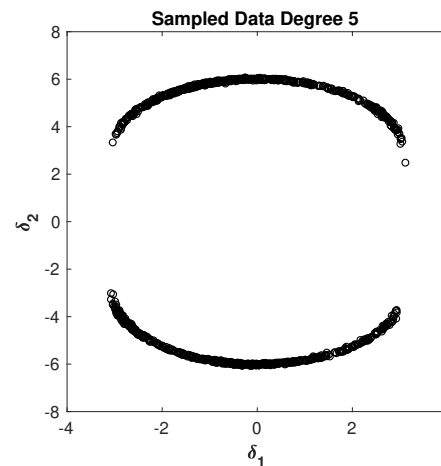
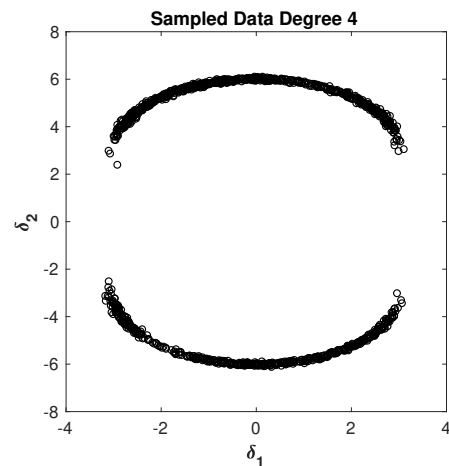
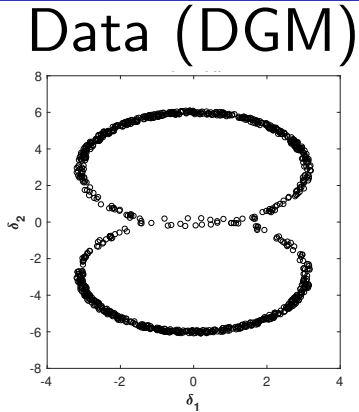
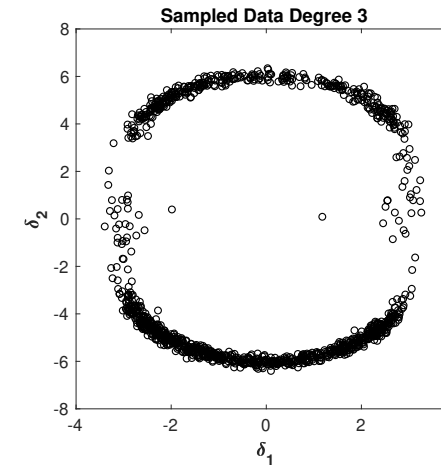
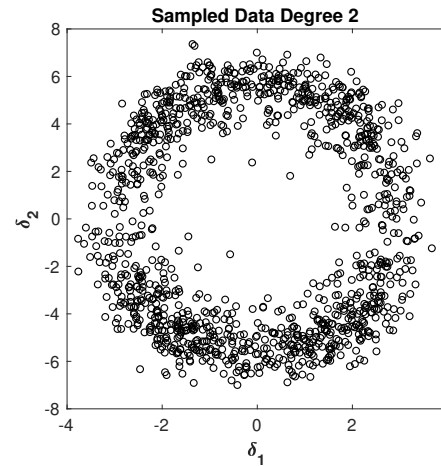
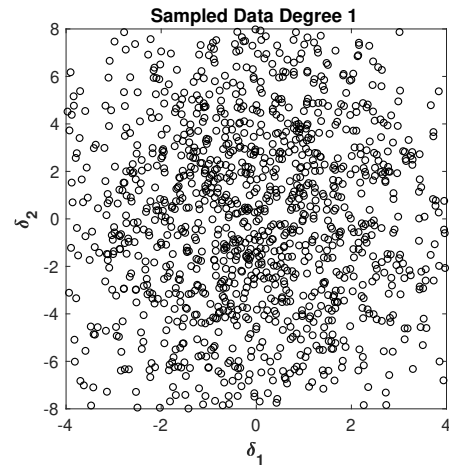
Example: Failure of the Lifted Approach (Prior Work)



- Problem: DGM is not modeled well using the sub-optimal hyperparameters.
- Best likelihood of the data is using a degree 3 SN.
- None of the density functions fit the data well.



Example: Failure of the Lifted Approach (Prior Work)

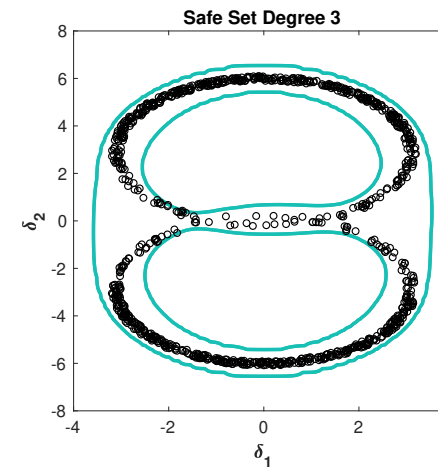
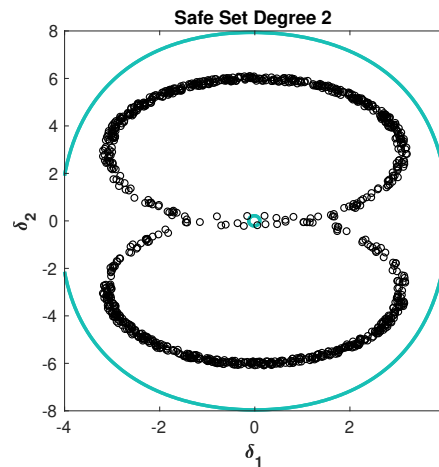
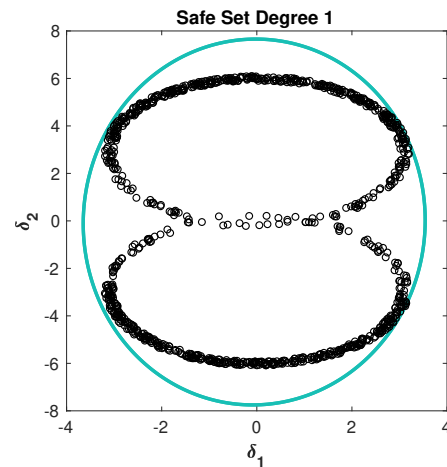


Bad: PDF generates no samples in regions with a lower density of points.

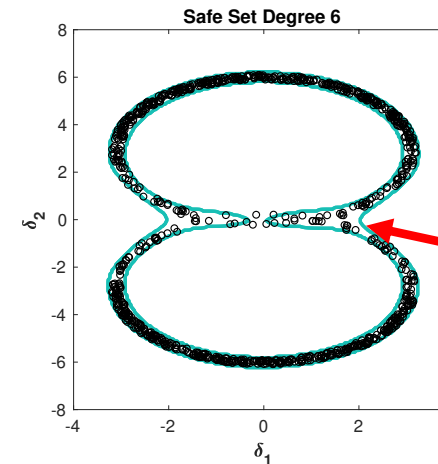
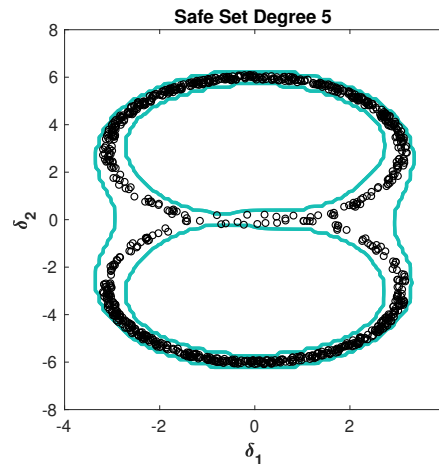
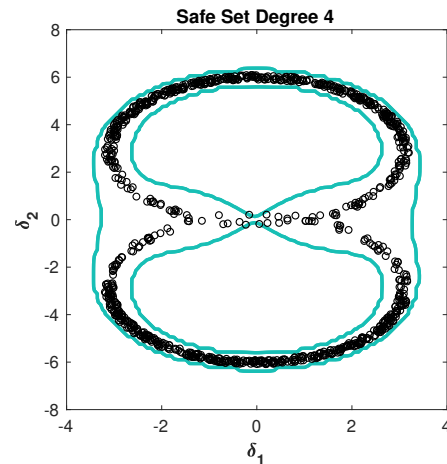
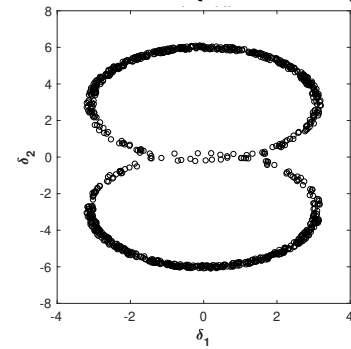
- Problem: DGM is not modeled well using the sub-optimal hyperparameters.
- Best likelihood of the data is using a degree 3 SN.
- Samples of the SN PDF don't match the data.



Example: Failure Success of the Lifted Level Sets (Prior Work)



Data (DGM)



Good: Scaled suboptimal level sets of PDF tightly enclose all the data points.

- Problem: DGM is not modeled well using the sub-optimal hyperparameters.
- Best likelihood of the data is using a degree 3 SN.
- A level set can be selected to fit the data “tightly” (improves with degree).

Conclusion: Scaling the PDF may help?



Improving the Likelihood of the Data for a SN

Goal: Modify the suboptimal hyperparameters to increase the likelihood of the data.

Improving Sliced Normal Hyperparameters

We will “rescale” P^* found in the “lifted space”.

$$\max_{\gamma \in \mathbb{R}^+} \left\{ \prod_{i=1}^m \frac{e^{-\phi(\delta^{(i)}; \mu^*, \gamma P^*, d)}}{\int_{\Delta} e^{-\phi(x; \mu^*, \gamma P^*, d)} dx} \right\} \quad (3)$$

Therefore:

$$\max_{\substack{\mu \in \mathbb{R}^q \\ P \in \mathbb{R}^{q \times q}}} \underbrace{L_f(\cdot; \mu, P, d)(\mathcal{D})}_{\text{Non-convex}} \geq \underbrace{L_f(\cdot; \mu^*, \gamma P^*, d)(\mathcal{D})}_{\text{Our Convex Method}} \geq \underbrace{L_f(\cdot; \mu^*, P^*, d)(\mathcal{D})}_{\text{Prior Work}}.$$

- We want the likelihood of the data $\mathcal{D}$ to be large.
- Our convex method scales P^* to optimize the likelihood of the data $\mathcal{D}$ not $\mathcal{D}_z$.
- Optimization Problem (3) has only one scalar decision variable γ .
- To solve Optimization Problem (3) we need to calculate the normalization constant.



Calculating the Normalization Constant

Monte Carlo Integration

Given a set of uniform samples over Δ , $\{s^{(i)}\}_{i=1}^{m_s}$, we can estimate the integration constant.

$$\int_{\Delta} e^{-\frac{(Z_d(s^{(i)}) - \mu^*)^T (\gamma P^*) (Z_d(s^{(i)}) - \mu^*)}{2}} d\delta \approx \hat{c}(\gamma) = \frac{1}{m_s} \sum_{i=1}^{m_s} e^{-\frac{(Z_d(s^{(i)}) - \mu^*)^T (\gamma P^*) (Z_d(s^{(i)}) - \mu^*)}{2}}$$

- The error of Monte Carlo integration decreases as $\frac{1}{\sqrt{m_s}}$.
- The decrease in error does not depend on the dimension of the integral.
- Even with millions of samples $\hat{c}(\gamma)$ can be calculated quickly.
- Care must be taken to select an m_s that is large enough.



Improving the Likelihood of the Data with Convex Optimization

Improving SN hyperparameters with Convex Optimization

We will optimize the log likelihood of the Sliced Normal.

$$\gamma^* = \operatorname{argmax}_{\gamma \in \mathbb{R}^+} \left\{ m \log(\hat{c}(\gamma)) - \frac{1}{2} \sum_{i=1}^m (Z_d(\delta) - \mu^*)^T (\gamma P^*) (Z_d(\delta) - \mu^*) \right\}$$

Where $\hat{c}(\gamma)$ is the numerical approximation of the normalization constant.

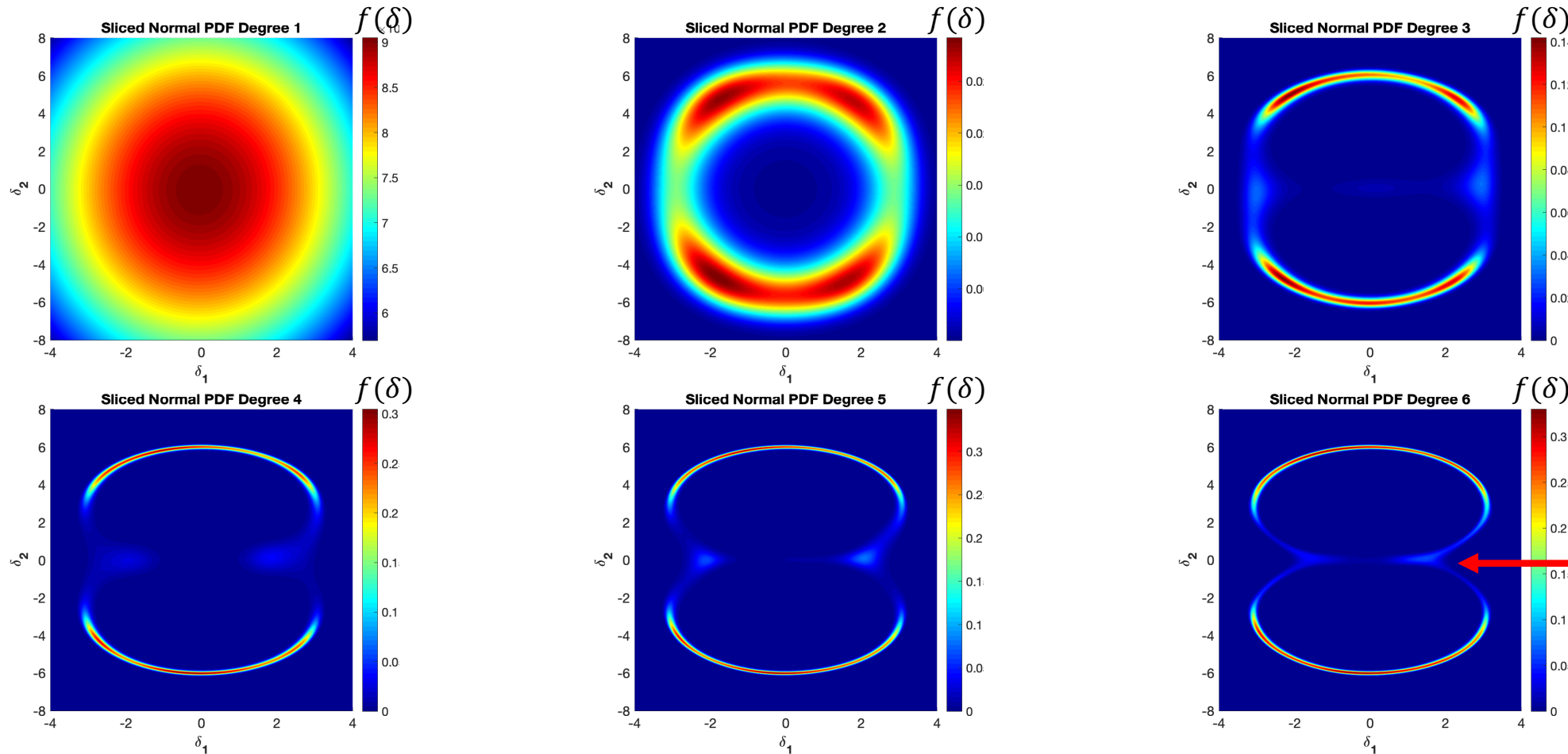
- This optimization problem is convex with respect to γ !
- We will use MATLABs `fmincon`.

The resulting Sliced Normal is:

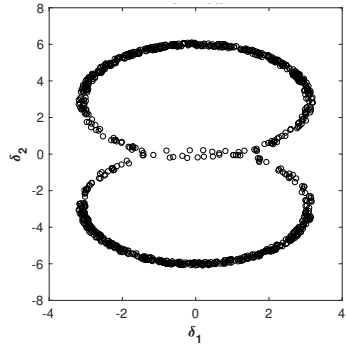
$$f(\delta; \mu^*, \gamma^* P^*, d) = \frac{1}{\hat{c}(\gamma^*)} e^{-\frac{1}{2} (Z_d(\delta) - \mu^*)^T (\gamma^* P^*) (Z_d(\delta) - \mu^*)}$$



Example: Improving the SN Hyperparameters



Data (DGM)

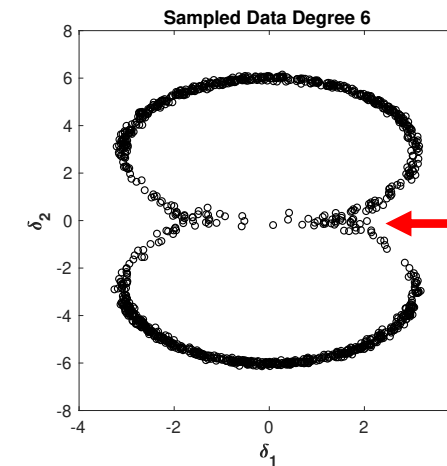
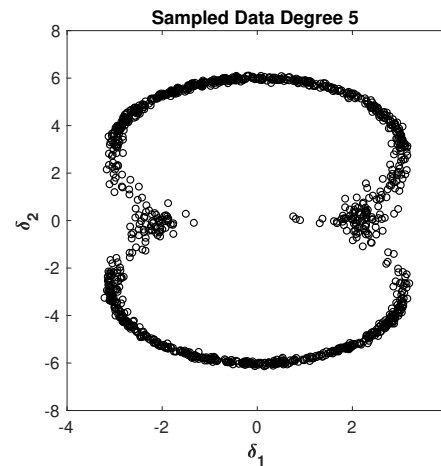
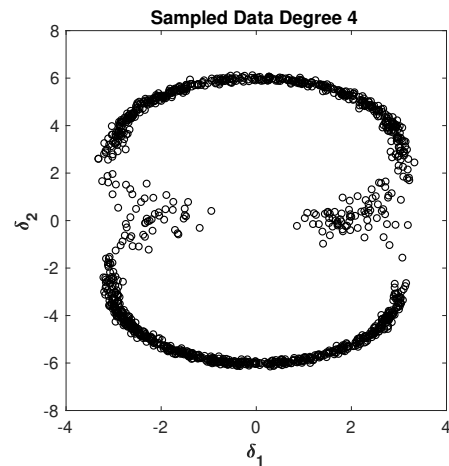
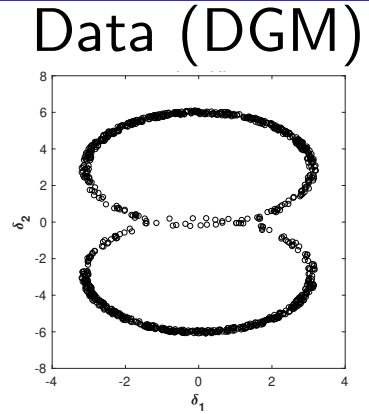
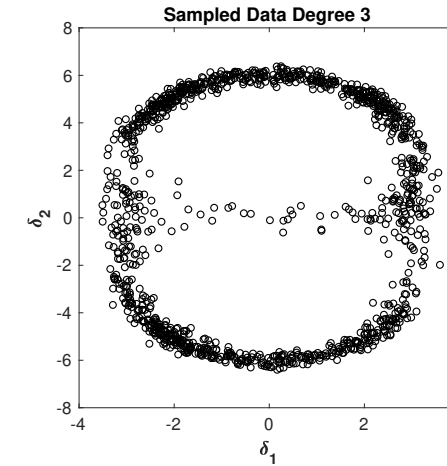
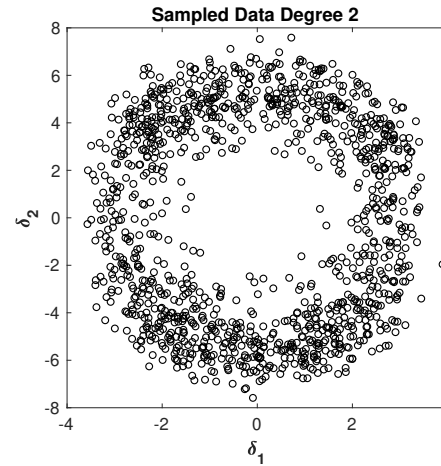
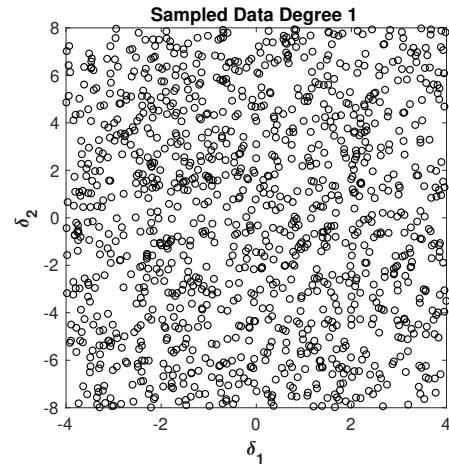


Good: PDF is non-zero in areas with a low density of data points.

- The likelihood of the data increases as the degree increases!
- The density functions fit the data better as the degree increases.



Example: Improving the SN Hyperparameters

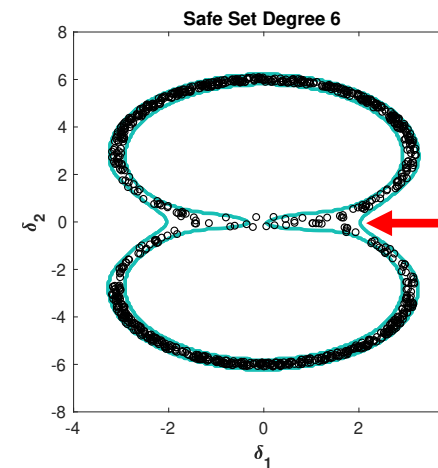
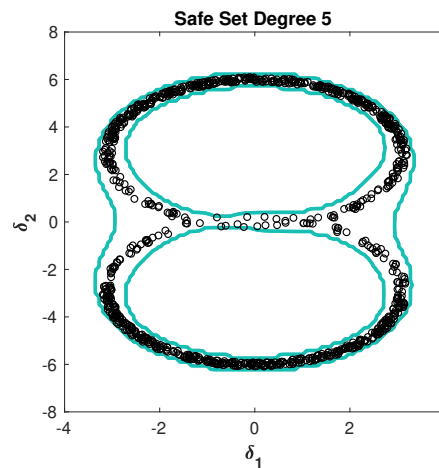
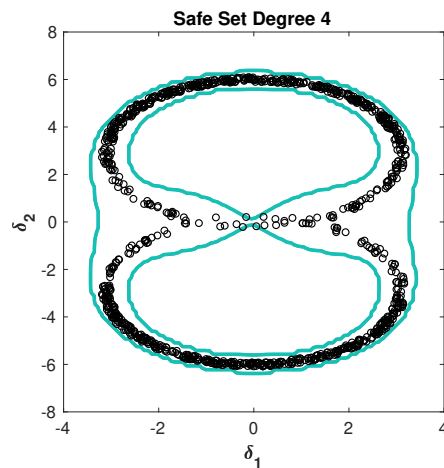
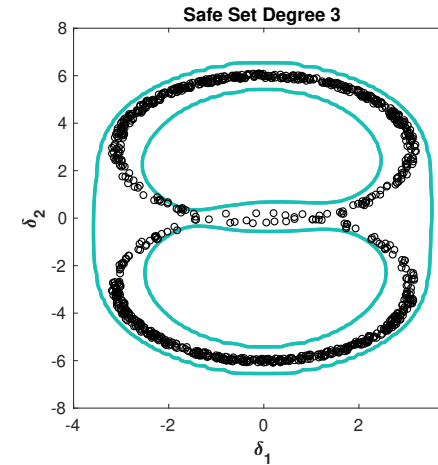
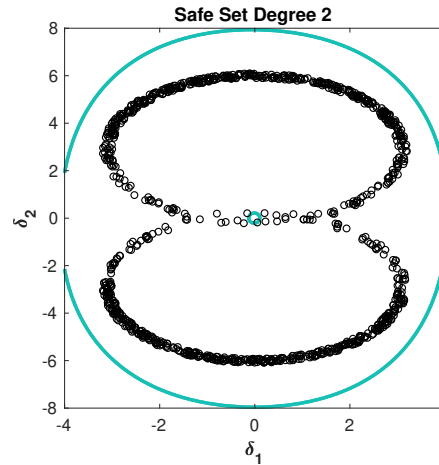
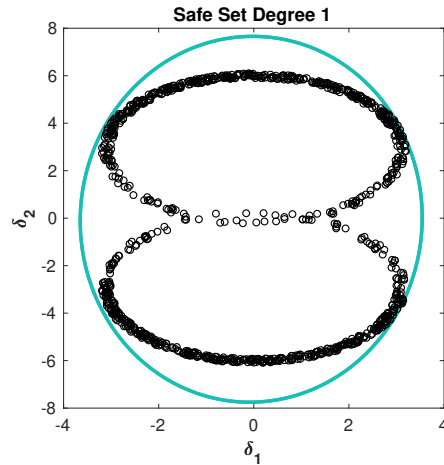


A few samples are in regions with a low density of points!

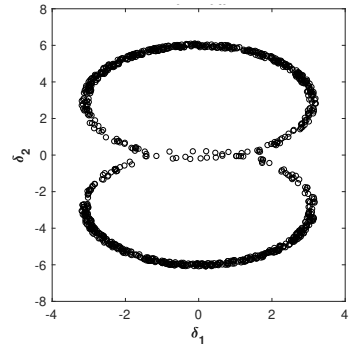
- The likelihood of the data increases as the degree increases!
- Samples begin to match the data (as degree increases).



Example: Improving the SN Hyperparameters



Data (DGM)

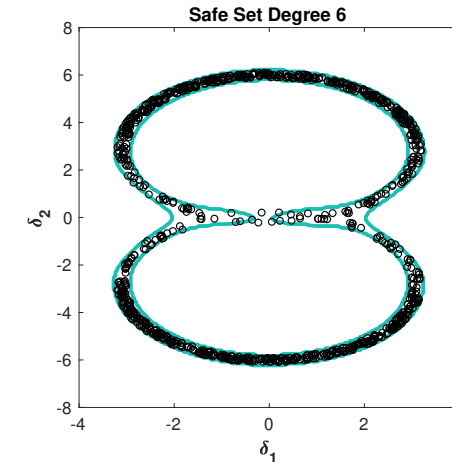
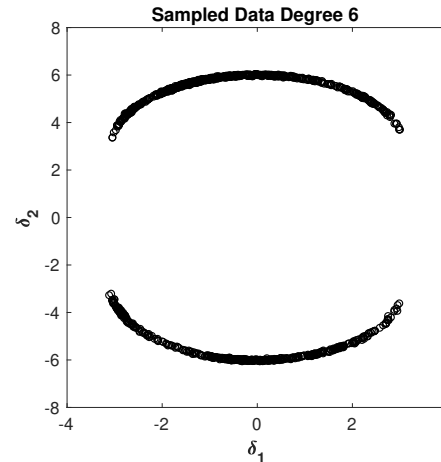
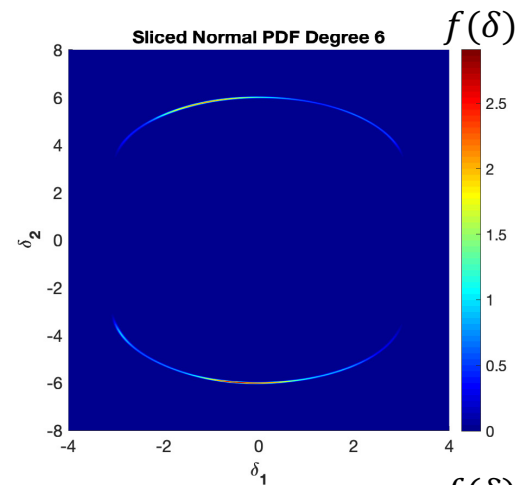


A level set still tightly encloses all the data points.

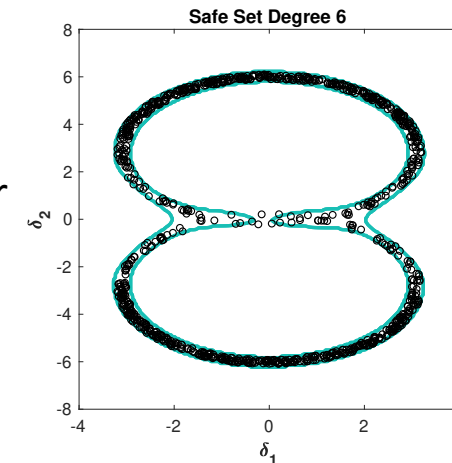
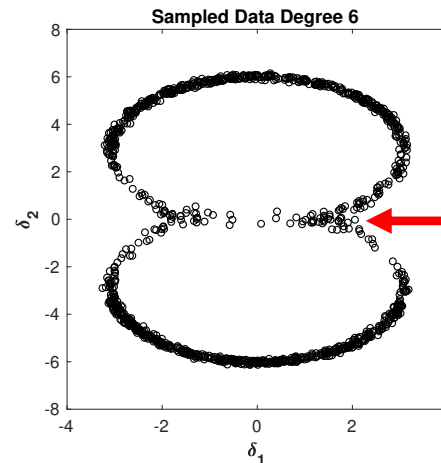
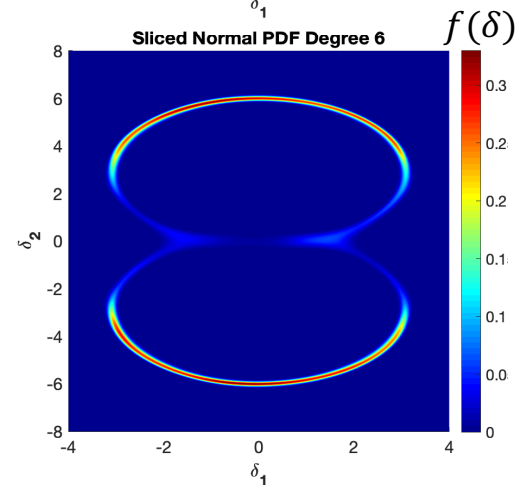
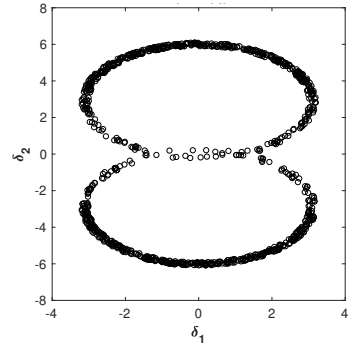
- The likelihood of the data increases as the degree increases!
- The data enclosing sets are the same!



Comparison Between Prior Method and Our Method (Degree 6 SN)



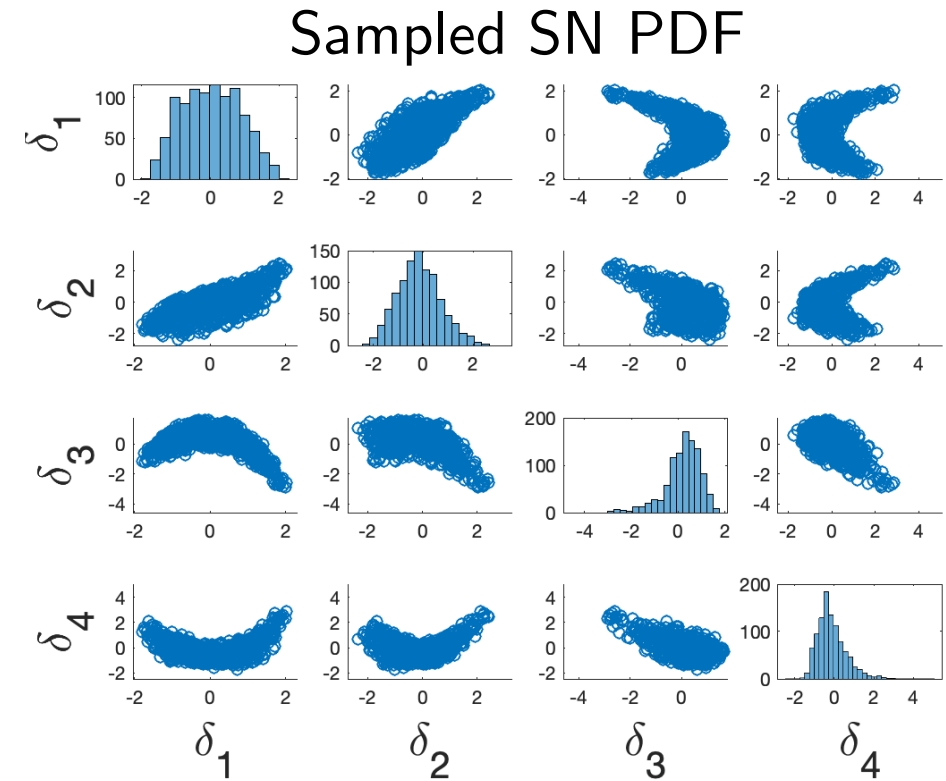
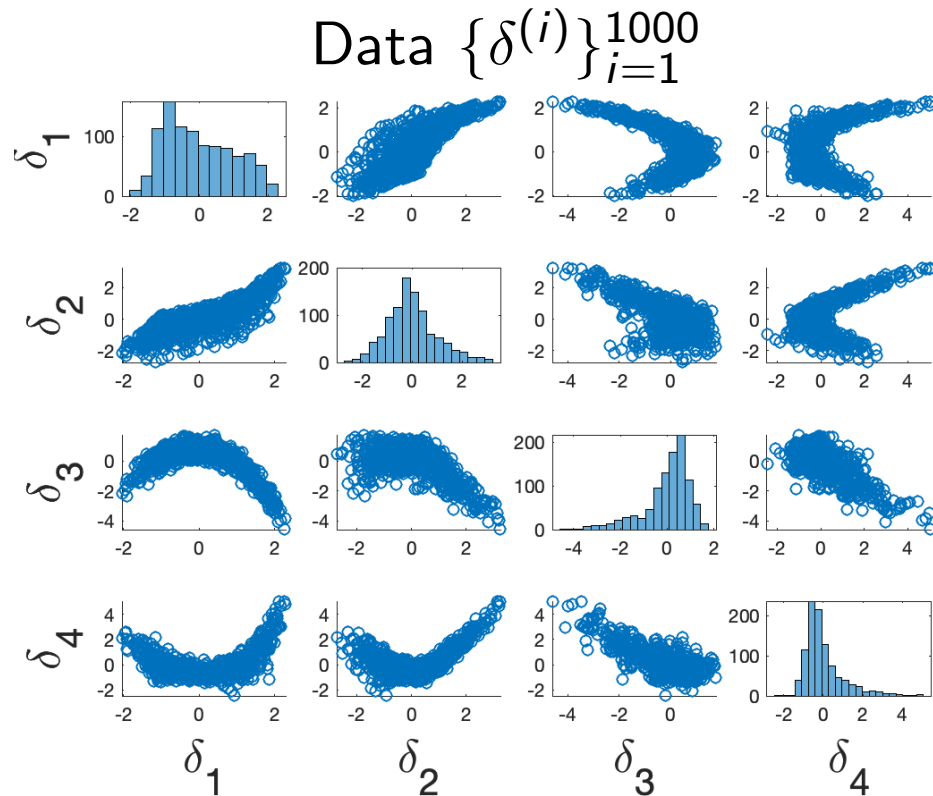
Data (DGM)



- Our method has a 69.6% increase in the likelihood of the data.



Example: Sampling Higher Dimensional Distributions



Example:

- We simulate a four dimensional distribution with strong parameter dependencies.
- We make no assumption on the underlying distribution of the data.



Conclusions

Given data, $\{\delta^{(i)}\}_{i=1}^m$ from a DGM of unknown structure, we can:

- Use Sliced Normal PDFs to model the DGM.
- Simulate new samples from the DGM.
- Find polynomial inequalities that enclose the data.

Important Properties of using this method:

- The analyst makes **no assumptions** on the structure of the DGM.
- Optimizing the hyperparameters involves solving a convex optimization problem with one scalar decision variable regardless of the dimension of δ .
- Increasing the degree generally increases the likelihood function value of the distribution.

