INFORMATION GEOEMTRY

DEFINITIONS

- Two important and related concepts have emerged
  1. **Practical Identifiability**
  2. **Sloppiness**
- The eigenvalues of the Fisher Information Matrix seem relevant to each.
- Can we now give a more rigorous definition?
  - How small does an eigenvalue need to be to be practically unidentifiable?
  - How much do the eigenvalues need to spread to be sloppy?
- Eigenvalues of FIM are problematic.

FITTING POLYNOMIALS

Example: Fitting polynomials by least squares on [0,1].

Approach 1: $y = \sum_n \theta_n t^n$
$I_{\mu\nu} = 2/(1 + \mu + \nu)$ is the Hilbert Matrix

Approach 2: $y = \sum_n \phi_n L_n(t)$ where $L_n(t)$ is the appropriately shifted Legendre polynomial.
$I_{\mu\nu} = \delta_{\mu\nu}$ is the identity matrix

Poll: Are these the same model?

PARAMETERIZATION DEPENDENCE

Given two parameterizations of a model, θ and ϕ, the FIM for the two parameterizations are related by:

$$\mathcal{I}_\theta = \left(\frac{\partial\phi}{\partial\theta}\right)^T \mathcal{I}_\phi \left(\frac{\partial\phi}{\partial\theta}\right)$$

$\mathcal{I}$ transforms like a covariant rank-2 tensor under reparameterization.

With an appropriate reparameterization, $\mathcal{I}$, can be transformed into any positive (semi-)definite matrix.

PARAMETERIZATION DEPENDENCE

Possibilities:

1. Practical Unidentifiability/Sloppines are consequences of poorly chosen parameters.
   They are not properties intrinsic to the model.
   - Why does sloppiness appear to be so ubiquitous?
   - Are we really that bad at modeling?
2. There is some other parameterization-invariant characterization.
   - Invariance to reparameterization sounds like a geometry problem.

INFORMATION GEOMETRY

The Fisher Information has all the properties of a Riemannian metric:

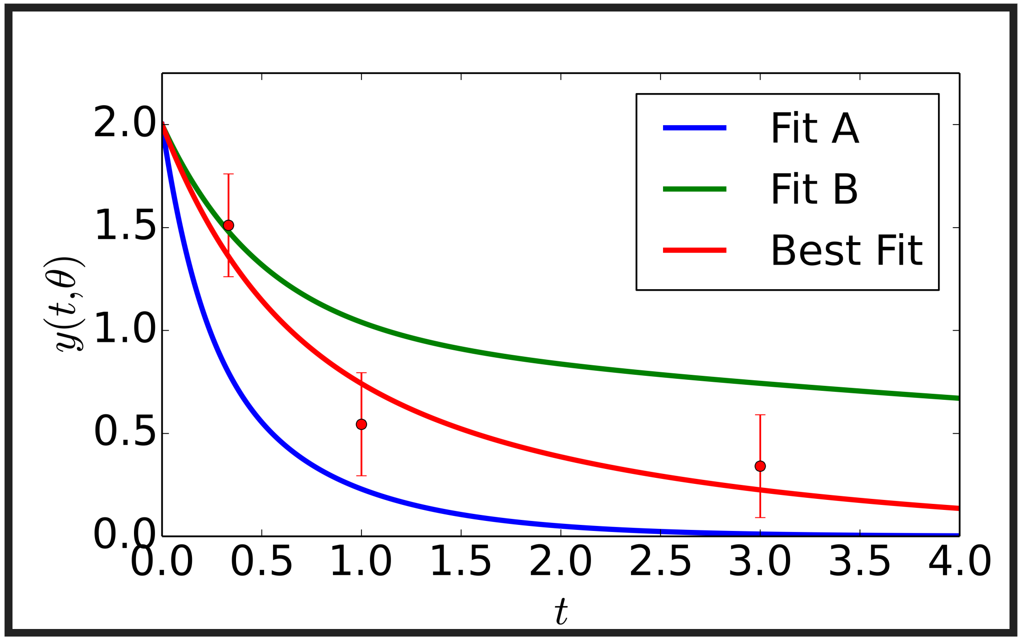
- Positive semi-definite
- Transforms like a covariant rank-2 tensor

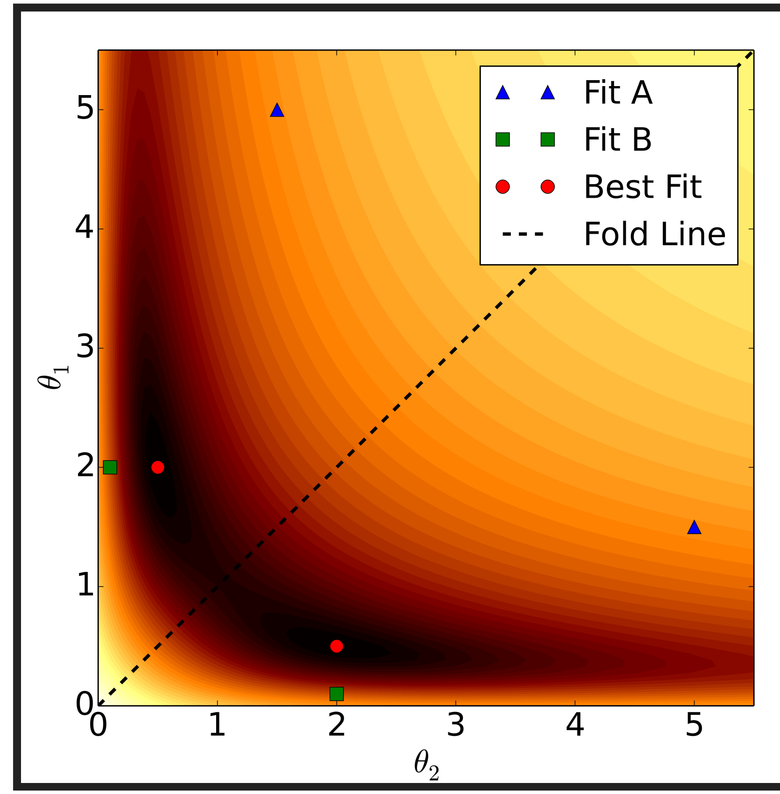
Let's take this interpretation literally. Perhaps there is a geometric insight (i.e., parameterization invariant) into why some models are unidentifiable and sloppy.

Our approach: Computational differential geometry using the FIM as the metric.

TWO EXPONENTIAL EXAMPLE:

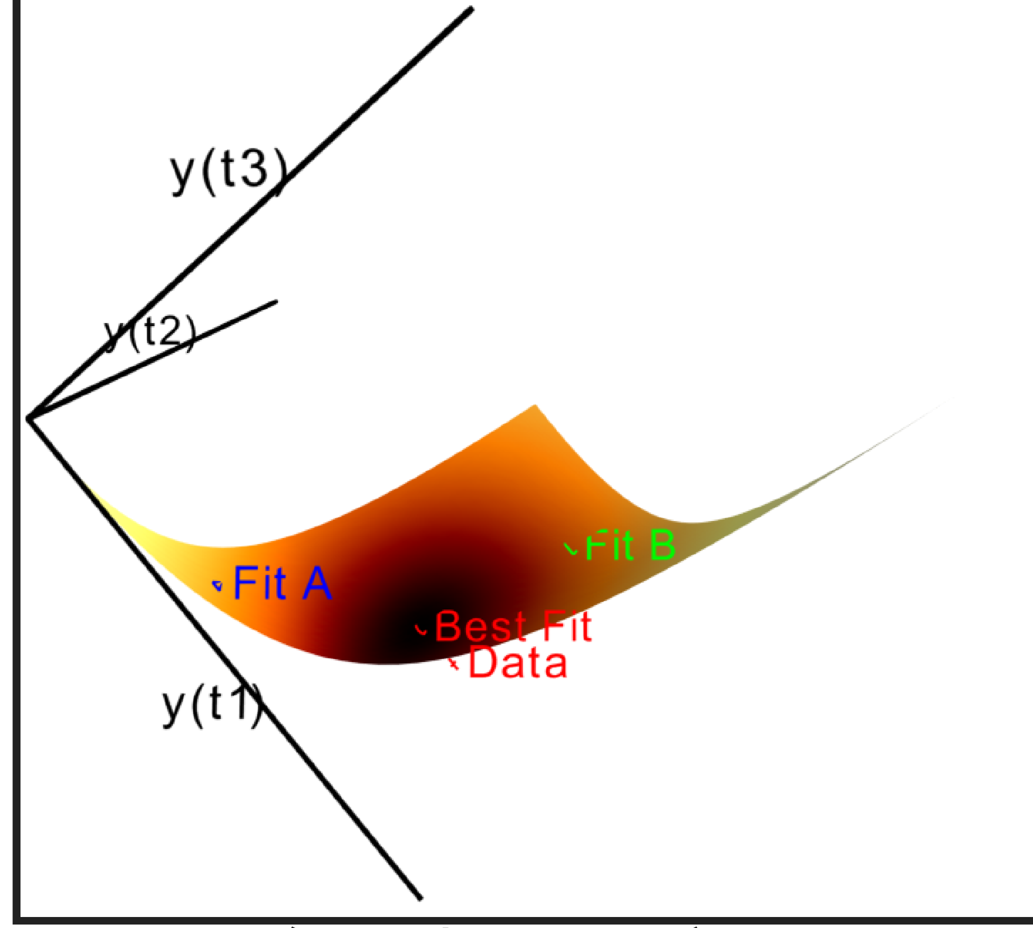
$$y(t,\theta) = e^{-\theta_1 t} + e^{-\theta_2 t}$$





DATA SPACE:

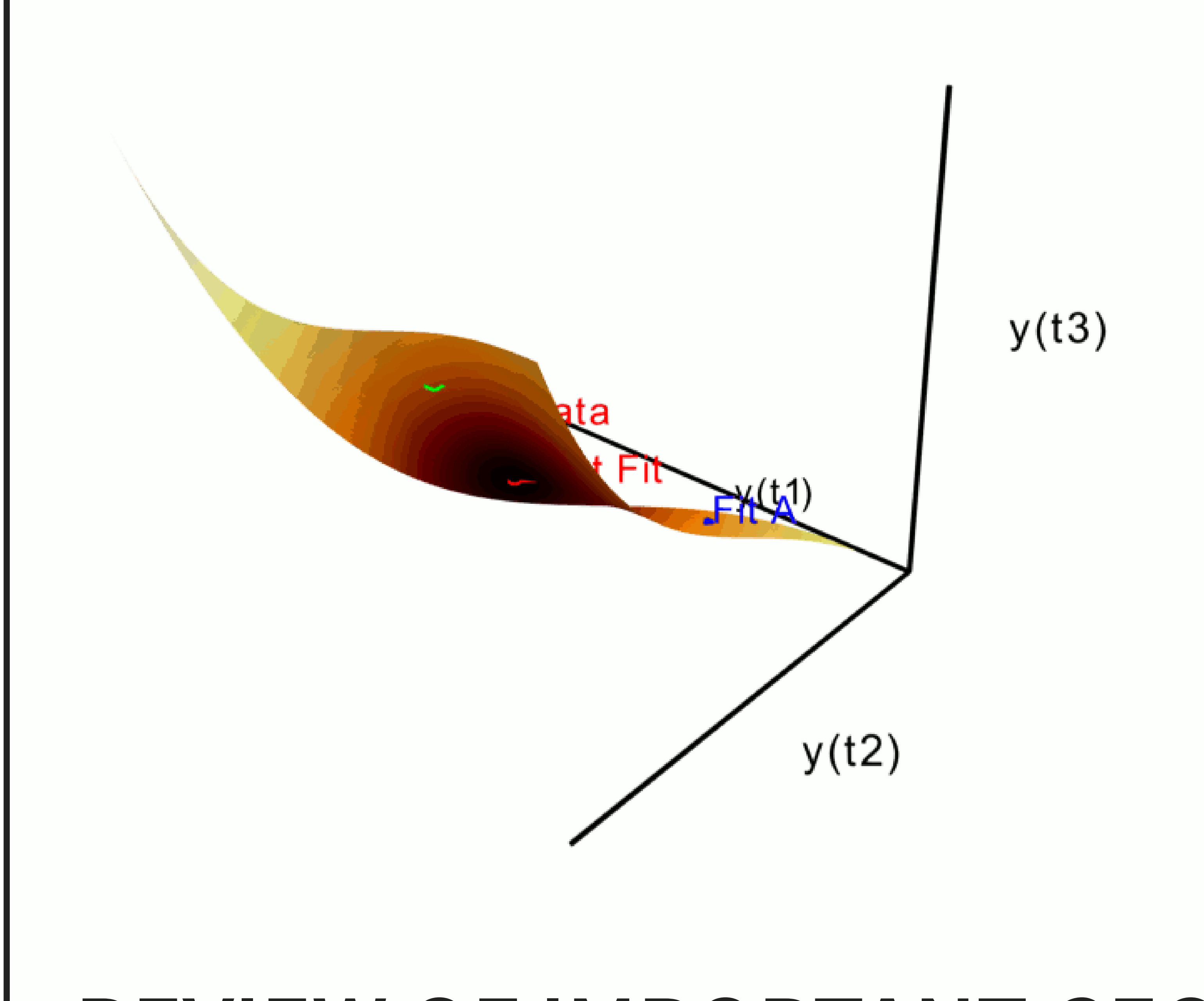
- One axis for each data point.
- Observed data becomes a vector $d_i \rightarrow \vec{\mathbf{d}}$
- Model Predictions become a vector $y_i(\theta) \rightarrow \vec{\mathbf{y}}(\theta)$
- Varying the parameters, sweeps out a surface: the *Model Manifold* $\mathcal{M}$



Quiz:

The dimensionality of the embedding space? (3 in this case?)
The number of data points

The dimensionality of the model manifold? (2 in this case?)
The number of locally structurally identifiable parameters



REVIEW OF IMPORTANT GEOMETRIC CONCEPTS

1. Embedding Space
2. Intrinsic vs. Extrinsic Properties
3. Geodesics
4. Curvature

EMBEDDING SPACE

- We can imagine the manifold living in (i.e., embedded in) a higher dimensional Euclidean space.
- The Euclidean inner product of the embedding space *induces* a metric on the manifold.

$$\mathbf{y}(\theta) \in \mathbb{R}^M, \theta \in \mathbb{R}^N$$

$$\mathbf{y}(\theta + d\theta) = \mathbf{y}(\theta) + d\mathbf{y} = \mathbf{y} + \frac{\partial \mathbf{y}}{\partial \theta} d\theta = \mathbf{y}(\theta) + J d\theta$$

$$dy^2 = d\mathbf{y} \cdot d\mathbf{y} = d\theta^T \left(J^T J \right) d\theta$$

- $J^T J = \mathcal{I} \equiv g$ is the metric on the tangent space.
- We refer to the embedding space as "data space" and denote it by $\mathcal{D}$.

LEAST SQUARES EMBEDDING

We have already seen in the toy example:

- One Euclidean embedding dimension for each residual.
- Distance is in units of standard deviations of the data. (Each data-space axis is $y_i(\theta)/\sigma_i$)

GENERAL EMEDDING

- For a general probability distribution, let $P_i(\theta)$ be the probability of the ith outcome. (i is a continuous index for probability densities)
- Let $z_i(\theta) = \sqrt{P_i}$, so that $\mathcal{M}$ is a subset of the hyper-sphere.
- **Exercise:** Show that a Euclidean distance in-z space induces the FIM as the metric on the tangent space.

RELATION BETWEEN EMBEDDINGS*

Exercise: Show that the distance function:

$$D(\theta_i, \theta_j) = -2 \log\left(\langle z(\theta_i), z(\theta_j) \rangle\right)$$

gives gives the Fisher Information Metric for infinitesimal distances.

Show that for the case of least-squares data fitting, this distance implies the least squares embedding.

*Katherine Quinn, unpublished.

INTRINSIC VS. EXTRINSIC

- In general, there many ways of isometrically embedding a particular manifold.

- Properties that depend on the embedding are called *extrinsic*.
- Properties independent of the embedding are called *intrinsic*.
- The metric, $\mathcal{I}$, is by definition intrinsic.
- Much of the foundational work in Information Geometry by Amari and others focuses on intrinsic properties.*
- Extrinsic properties are useful for statistics and pioneered by Bates and Watts.**
  - Observed data is off the manifold.
  - Cost = distance through embedding space to the data.
  - Extrinsic curvature $\implies$ local minima in cost surface.

*Amari, Shun-ichi, and Hiroshi Nagaoka. Methods of information geometry. Vol. 191. American Mathematical Soc., 2007.
**Bates, Douglas M. Watts, Donald G. Douglas M. Bates, and Donald G. Watts. Nonlinear regression analysis and lts applications. No. 519.536 B3. 1988.

VISUALIZATIONS

The high dimensionality of $\mathcal{D}$ and $\mathcal{M}$ make visualizations difficult.

One approach:

- Generate a sampling of points in parameter space.
  - Grid in parameter space
  - Sample geometrically motivated distributions (Ben Machta)
- Find the model predictions (vector) for each point and arrange them (mean shifted) in a matrix

$$\tilde{\mathbf{y}}_i = \mathbf{y}_i - \frac{1}{P}\sum_j \mathbf{y}_j$$

$$Y = \left[\tilde{\mathbf{y}}_1 \tilde{\mathbf{y}}_2 \ldots \tilde{\mathbf{y}}_P\right]$$

- Perform a PCA of these points:

$$Y = U\Sigma V^T$$

- Plot the first several PCA directions:

$$U\Sigma = YV$$

VISUALIZATION

Given the matrix of mean-shifted matrix of points:

$$Y = \left[\tilde{\mathbf{y}}_1 \tilde{\mathbf{y}}_2 \ldots \tilde{\mathbf{y}}_P\right]$$

we can also construct a projection matrix:

$$M = YY^T = U\Sigma^2 U^T$$

$$M_{ij} = \tilde{\mathbf{y}}_i \cdot \tilde{\mathbf{y}}_j$$

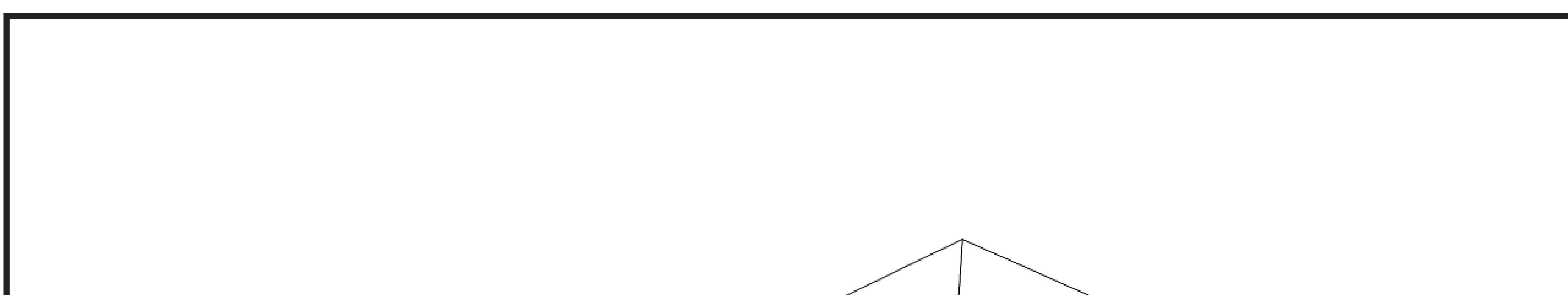
An eigenvalue decomposition of M is sufficient to produce an isometric embedding/visualization: $U\Sigma$.

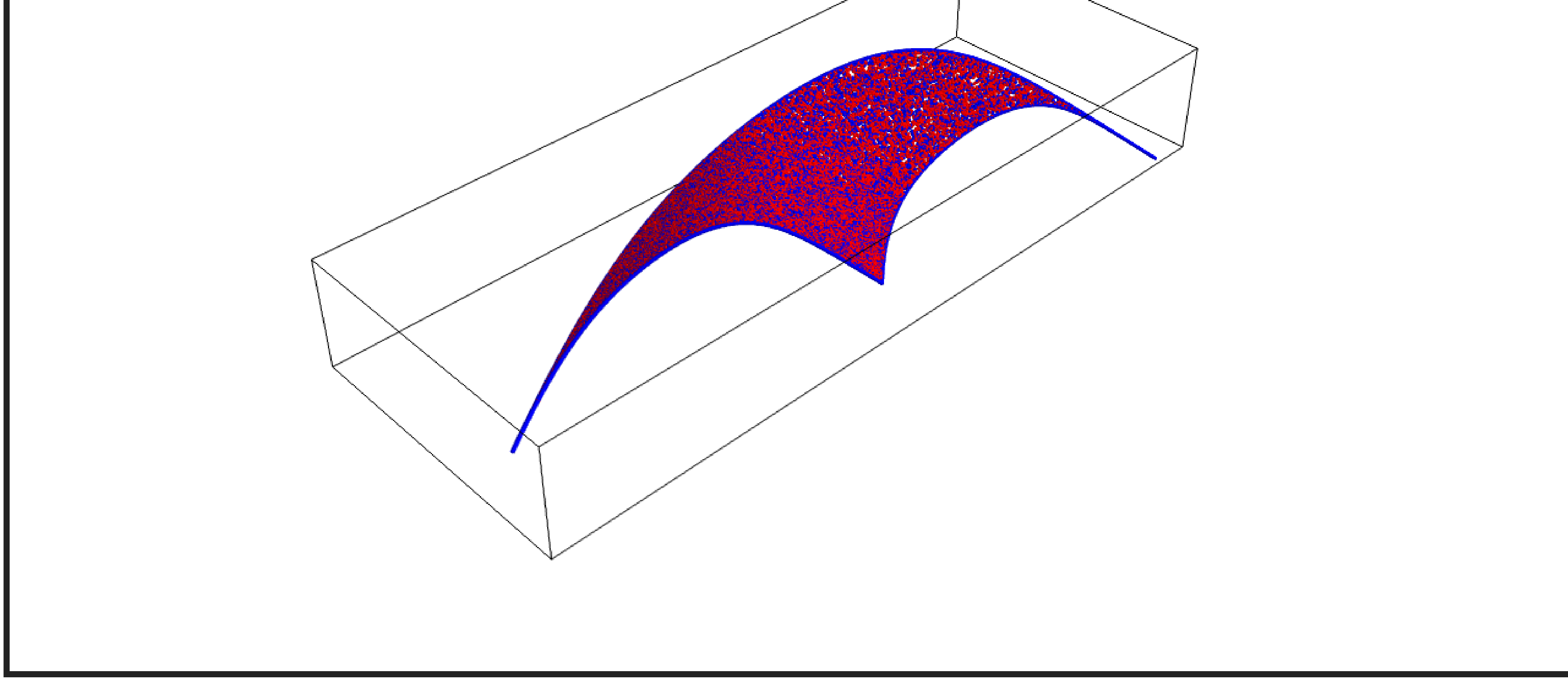
GALLERY OF MODEL MANIFOLDS

$$y = e^{-\theta_1 t} + e^{-\theta_2 t}$$

$N = 2$ Parameters

$M = 3$ Data points

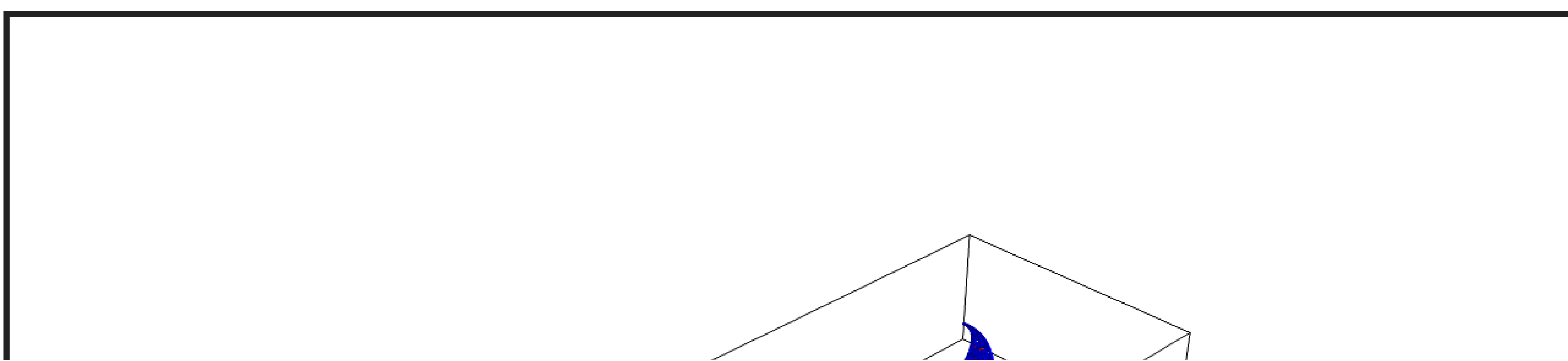


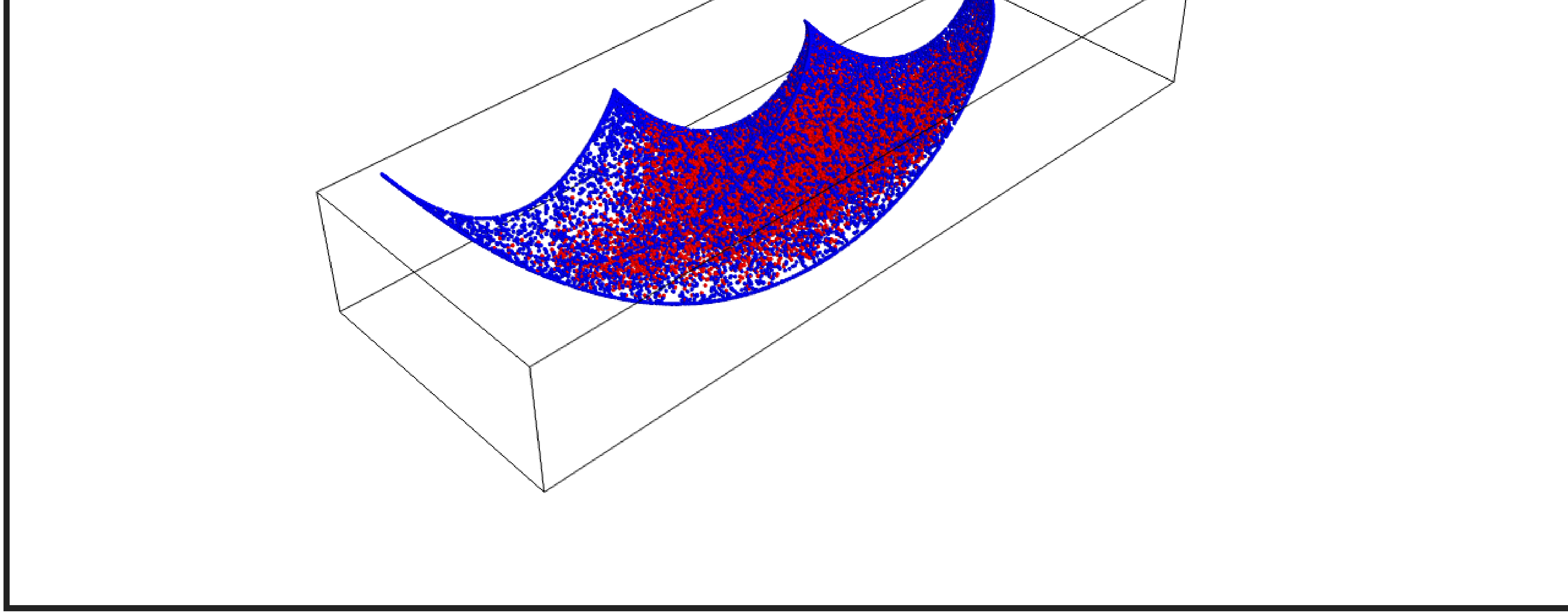


$y = e^{-\theta_1 t} + e^{-\theta_2 t} + e^{-\theta_3 t}$

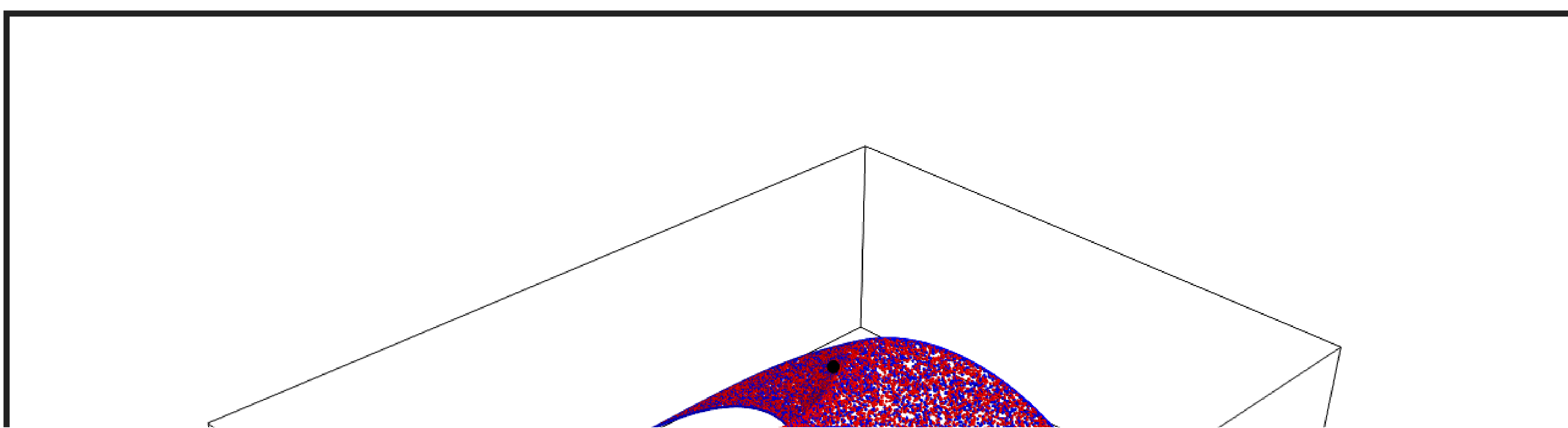
$N = 3$ Parameters

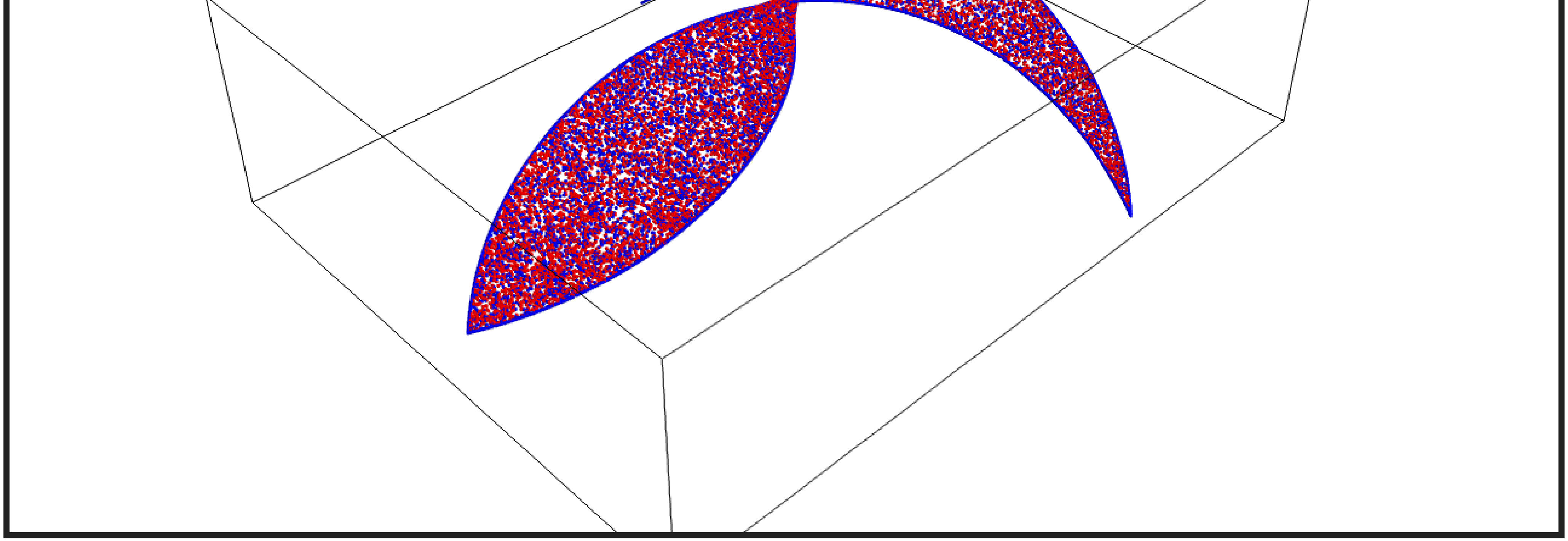
$M = 5$ Data points



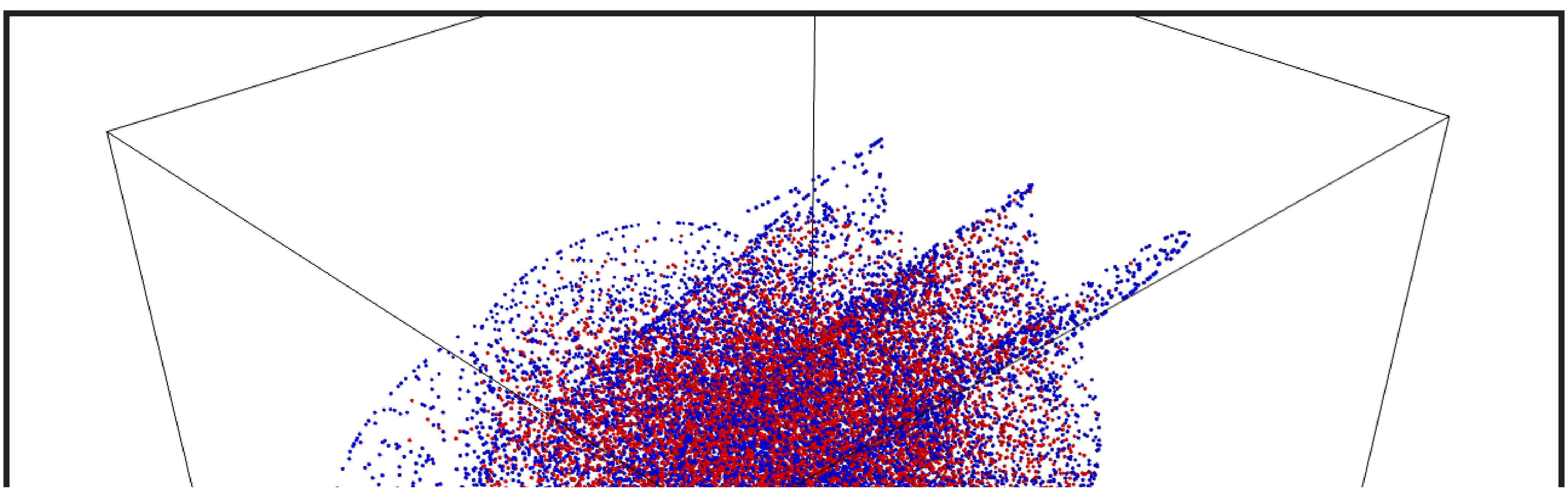


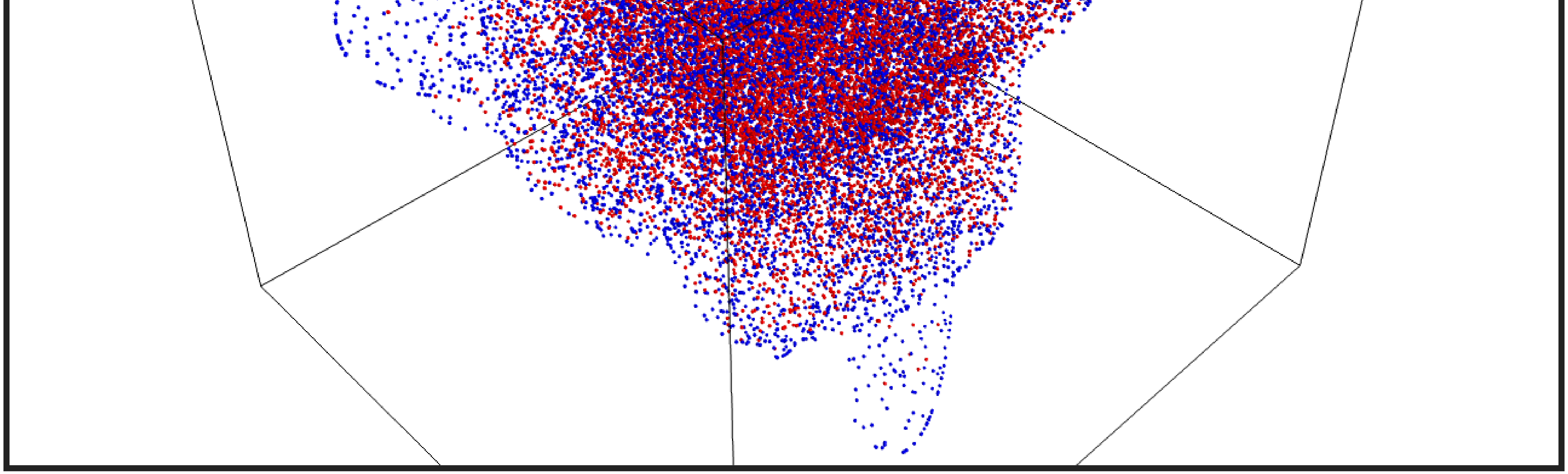
Enzyme Catalyzed Reaction (Minpack-2)
$N = 4$ (2 Dimensional Cross Section) Parameters
$M = 11$ Data points





Chebychev Quadrature (Minpack-2)
$N = 3$ Parameters
$M = 5$ Data points

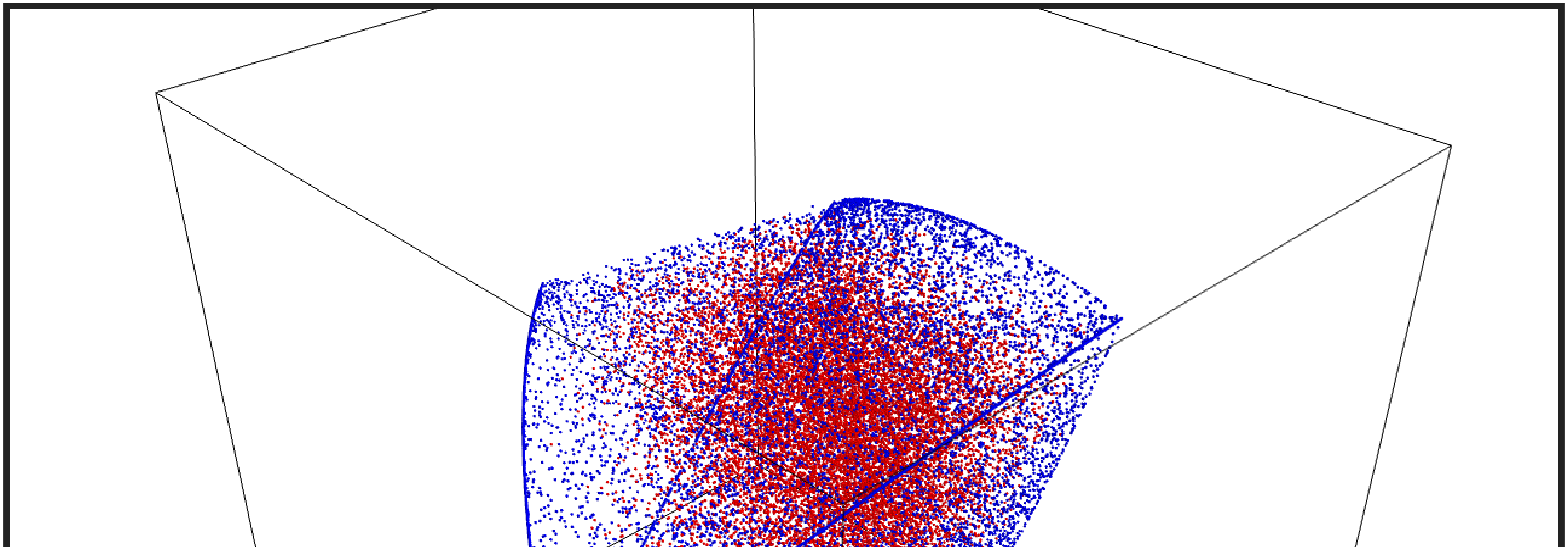


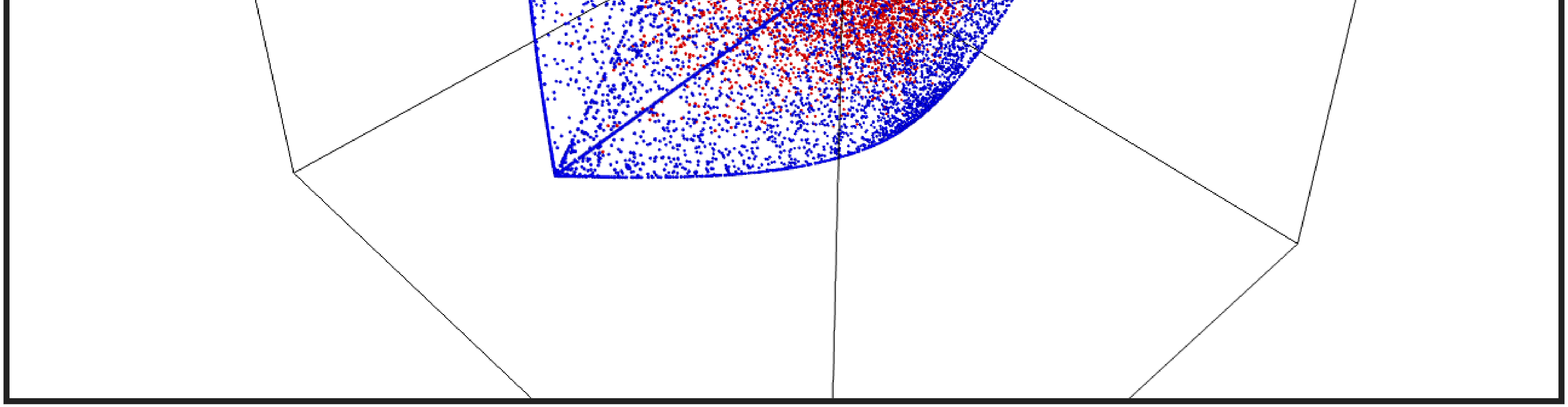


Isomerization of α-pinene (Minpack-2)

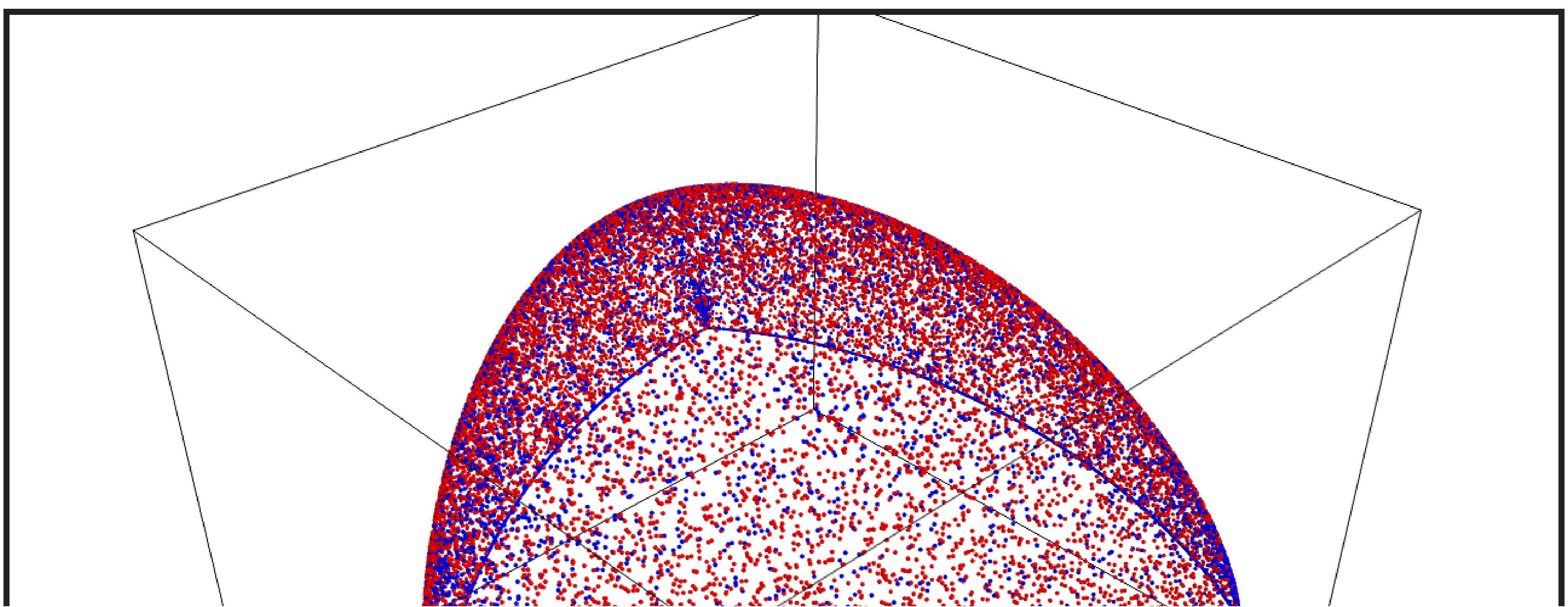
$N = 5$ Parameters

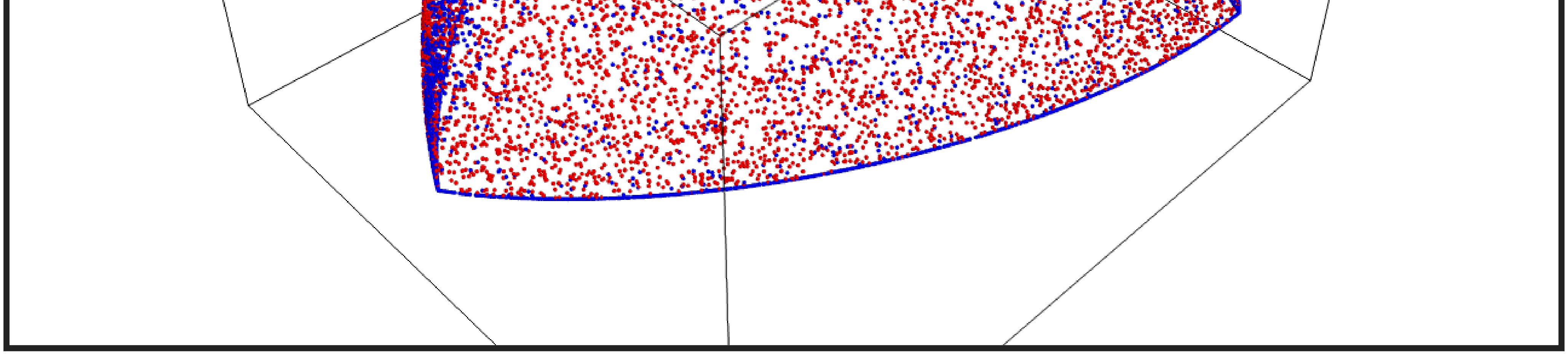
$M = 40$ Data points





2D Ising Model (2x2 Unit cell)
$N = 2$ Parameters (couplings only)
$M = 16$ "Data points" (16 distinct states)





GEODESICS

- Special paths on the model manifold.
- Satisfies a differential equation.
- Parallel transport of tangent vector.
  - Initial Value Problem
- Distance minimizing curves
  - (When using the metric

connection)

- Boundary Balue Problem

CURVATURE

Three types of curvature:

- Intrinsic (Riemann) Curvature
- Extrinsic Curvature
- Parameter-Effects Curvature

INTRINSIC VS. EXTRINSIC

- Intrinsic Curvature $\implies$ Extrinsic Curvature
  - Converse not true
- Ruled surfaces
  - Zero intrinsic curvature but nonzero extrinsic curvature
  - Example: Cylinder
- Large Extrinsic curvature associated with local minima of the cost

MEASURE OF EXTRINSIC CURVATURE

GEODESIC CURVATURE

SHAPE OPERATOR

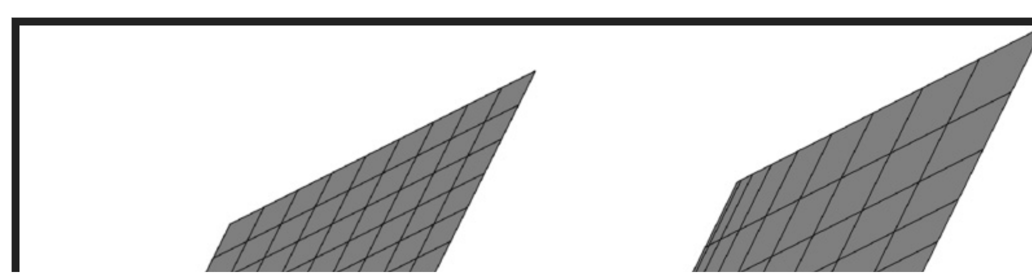
$J = \partial_\mu y = U \Sigma V^T, \quad P^N = 1 - UU^T, \quad S_{\mu\nu} = \hat{n} \cdot \partial_\mu \partial_\nu y$

$$\mathbf{v} = J\dot{\theta}, \mathbf{a} = P^N \partial_\mu \partial_\nu \mathbf{y} \dot{\theta}^\mu \dot{\theta}^\nu$$

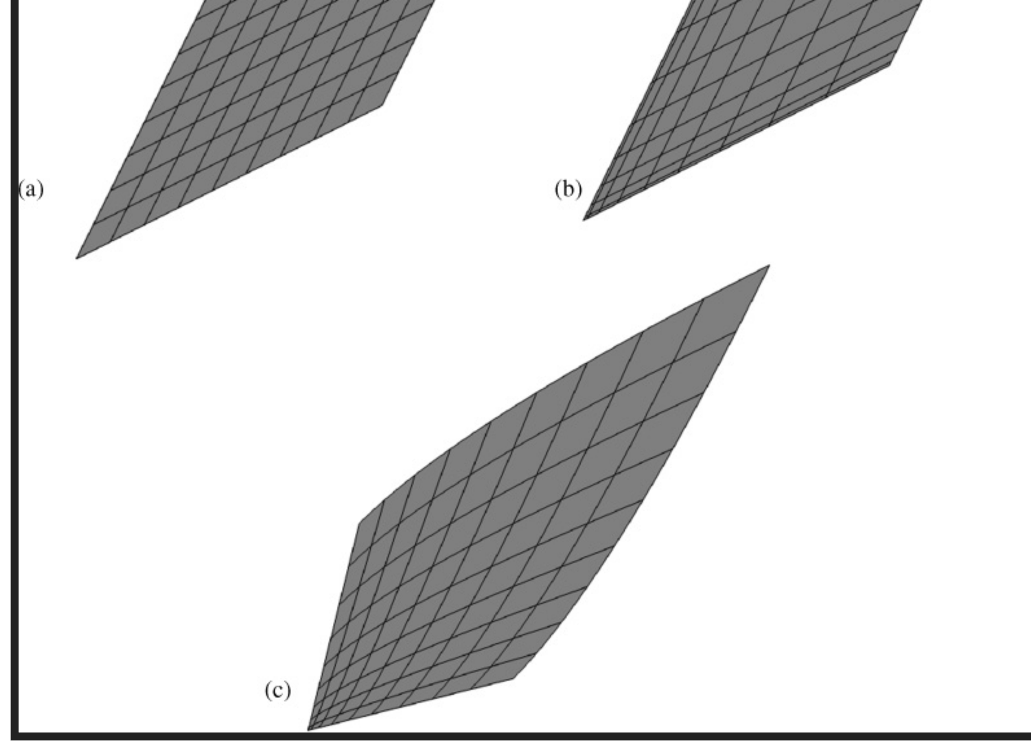
$$K = R^{-1} = \frac{|\mathbf{a}|}{|\mathbf{v}|^2}$$

Transtrum, Mark K., Benjamin B. Machta, and James P. Sethna. "Geometry of nonlinear least squares with applications to sloppy models and optimization." Physical Review E 83.3 (2011): 036701.

PARAMETER-EFFECTS CURVATURE



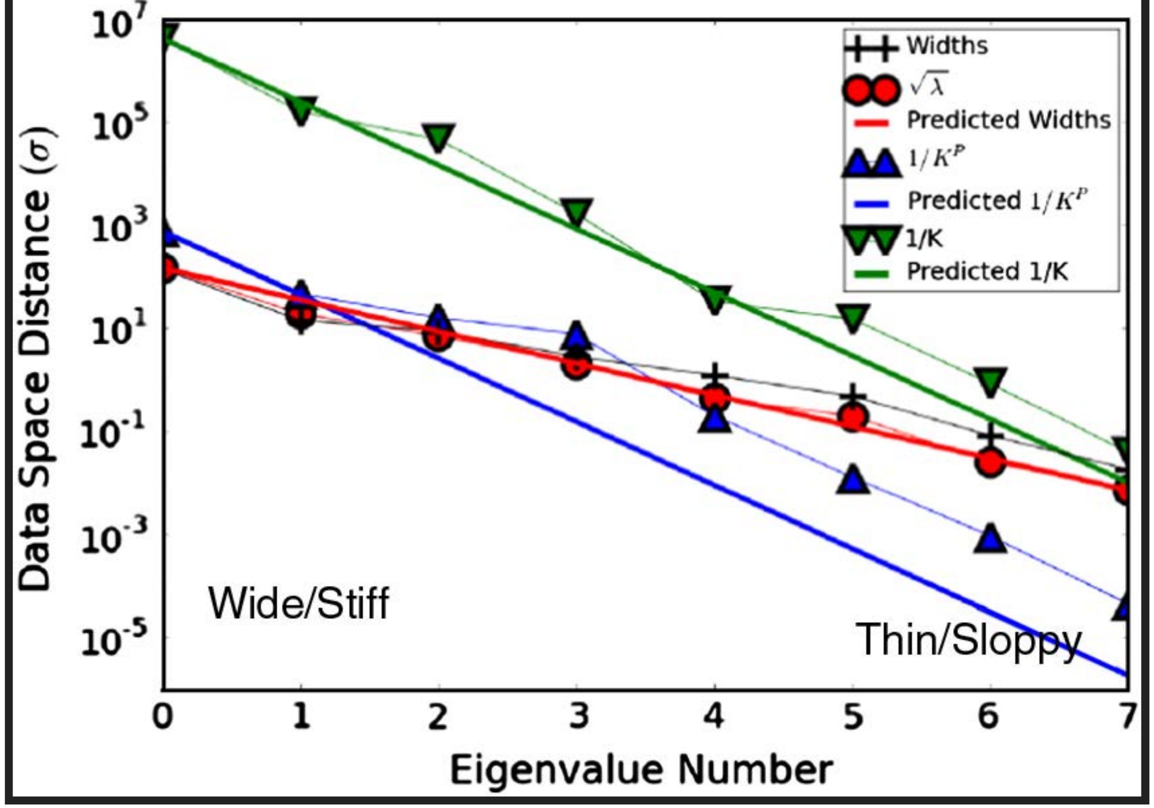
- Non-standard
- Introduced by Bates and Watts.*
- Bending/Stretching of the coordinate grid on the model manifold
- Same information as the connection coefficients
- In most cases it is much larger than either extrinsic or intrinsic curvatures



*Bates, Douglas M., and Donald G. Watts. "Relative curvature measures of nonlinearity." Journal of the Royal Statistical Society. Series B (Methodological) (1980): 1-25.
Transtrum, Mark K., Benjamin B. Machta, and James P. Sethna. "Geometry of nonlinear least squares with applications to sloppy models and optimization." Physical Review E 83.3 (2011): 036701.

GEOMETRIC SLOPPINESS: WIDTHS AND CURVATURES

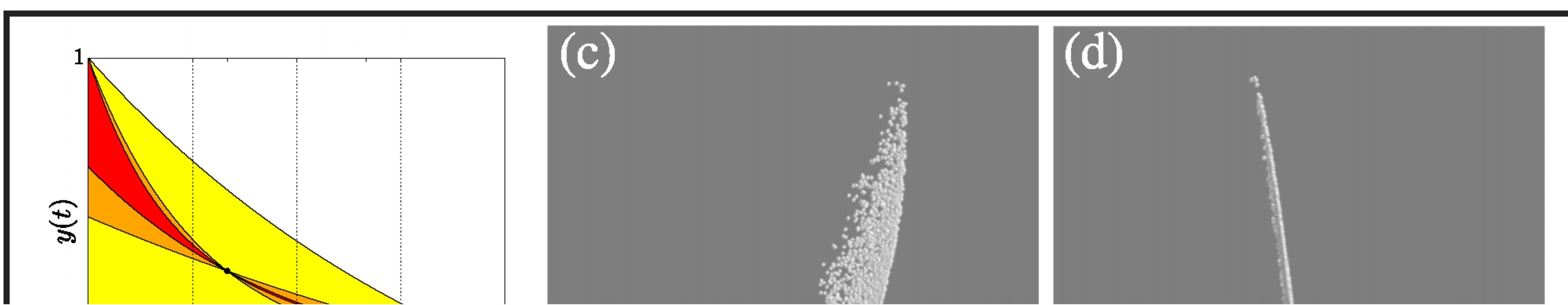
- Is there a parameterization-independent (geometric) characterization of sloppiness?

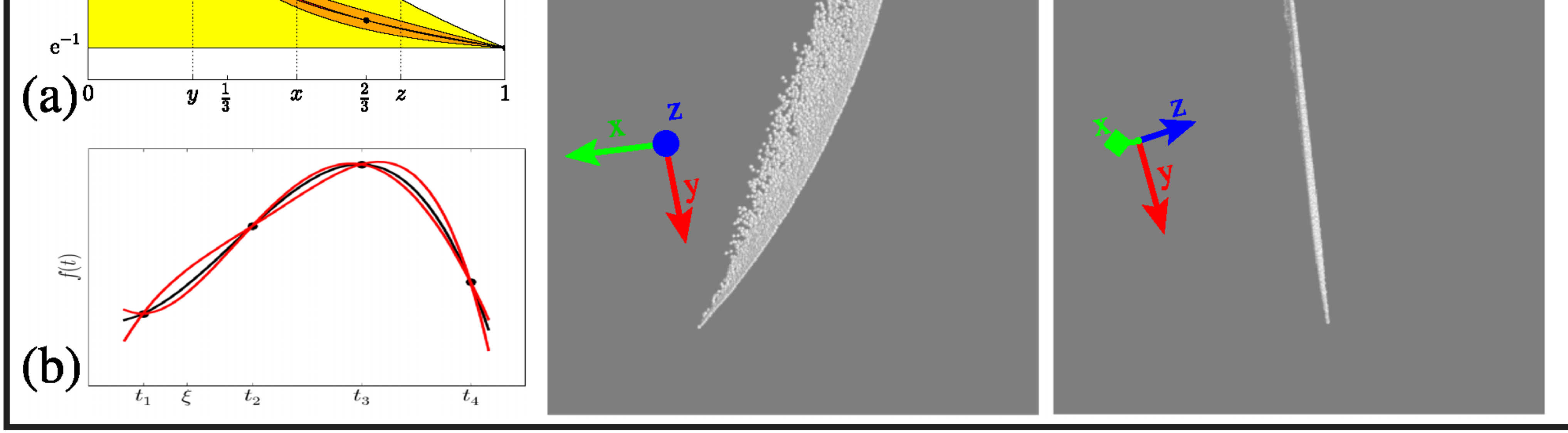


Transtrum, Mark K., Benjamin B. Machta, and James P. Sethna. "Why are nonlinear fits to data so challenging?." Physical review letters 104.6 (2010): 060201.

INTERPOLATION (PREVIEW)

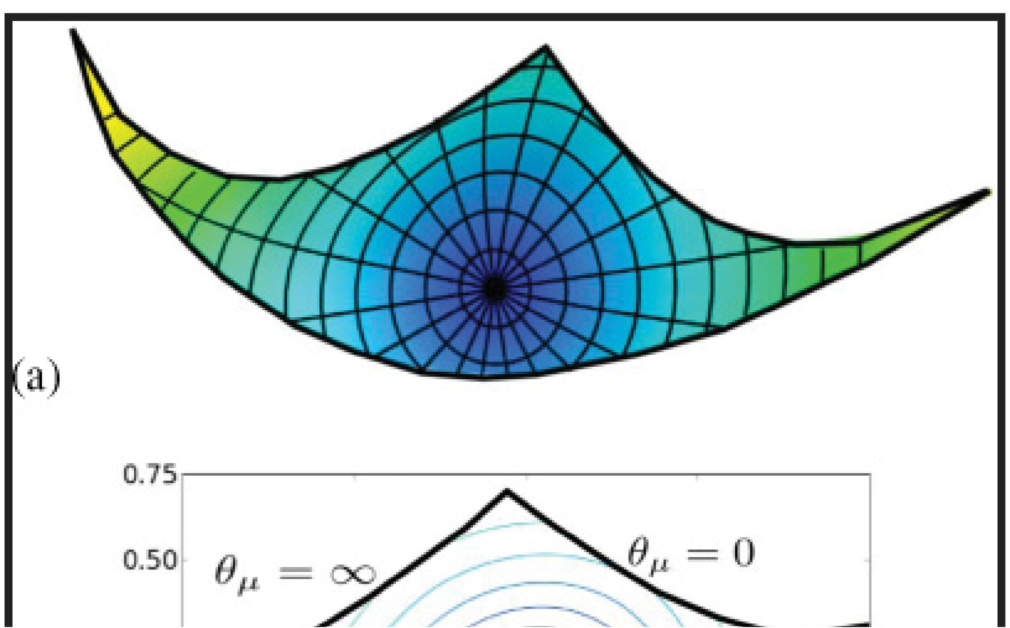
Why is the model manifold so thin?

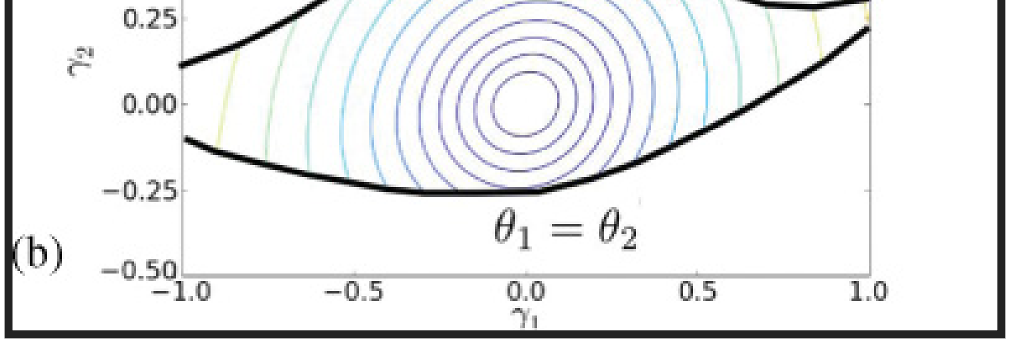




Transtrum, Mark K., et al. "Perspective: Sloppiness and emergent theories in physics, biology, and beyond." The Journal of chemical physics 143.1 (2015): 010901.

EXTENDED GEODESIC COORDINATES





- Use geodesics to construct new coordintes
- By construction: minimal parameter effects curvature