Geometric Foundation of Parameter Inference

Master's Colloquium

Rafael Arutjunjan

Chair of Theoretical Physics I, FAU
& Center for Astronomy Heidelberg

10. September 2020

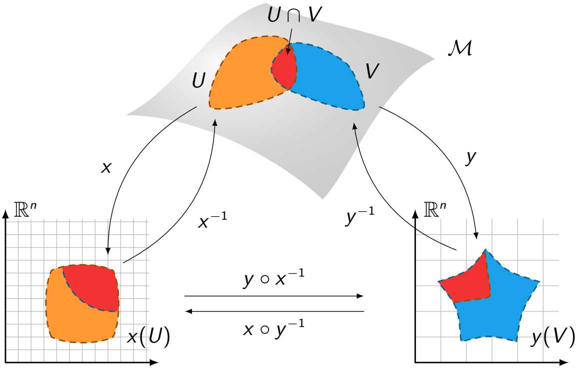
Table of Contents

Basics of Information Geometry

Parameter Inference and Information Geometry

- Parameter Inference: Given a dataset and a model, for which parameter configuration does the model most faithfully describe the data?
- What is the uncertainty in the "optimal" parameter configuration that was found?
- Information Geometry: Rephrase statistical problems in such a way that they can be given a geometric interpretation.
- Use powerful toolkit of differential geometry which focuses on intrinsic quantities which are invariant under reparametrisation.
- Almost no information-geometric literature on uncertainty and confidence regions available.

Chart Philosophy



- Coordinate representations give a potentially distorted view of the underlying "real system".
- Separate coordinate effects from intrinsic properties of the underlying manifold.

Information Divergences between Probability Distributions

- Quantify separation / dissimilarity between probability distributions e.g. using so-called information divergences, which are positive-definite functionals.
- For example, the Kullback–Leibler divergence D_{KL} defined by

$$D_{\mathrm{KL}}[p : q] := \int \mathrm{d}y\, p(y) \ln\left(\frac{p(y)}{q(y)}\right) = \mathbb{E}_p\Big(\ln\big(p/q\big)\Big). \tag{1}$$

- Can be interpreted as measuring loss of information (i.e. relative increase in Shannon entropy) by approximating a distribution $p(x)$ through $q(x)$.
- Advantages: invariant under reparametrisation, applicable between any two distributions with common support.
- Problem: typically not symmetric and does not satisfy a triangle inequality $\implies$ not a distance function.

Fisher Metric on Probability Spaces

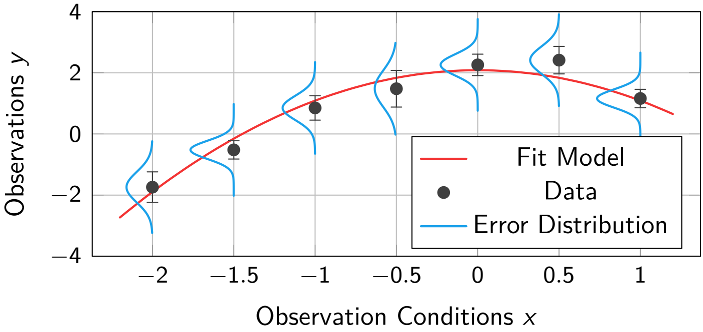
- Can establish so-called Fisher metric on spaces of probability distributions via Hessian of Kullback–Leibler divergence:

$$g_{ab}(\theta) := \left[\frac{\partial^2}{\partial\psi^a\,\partial\psi^b}\,D_{\mathrm{KL}}\Big[p(y;\theta) : p(y;\psi)\Big]\right]_{\psi=\theta} \tag{2}$$

$$= \ldots = -\mathbb{E}_p\left(\frac{\partial^2 \ln(p)}{\partial\theta^a\,\partial\theta^b}\right) = \ldots = \mathbb{E}_p\left(\frac{\partial \ln(p)}{\partial\theta^a}\,\frac{\partial \ln(p)}{\partial\theta^b}\right) \tag{3}$$

- Fulfils all the necessary requirements for a Riemannian metric tensor (symmetry, positive-definiteness, transformation behaviour).
- First employed by C. Rao in 1945 to study manifolds of probability distributions.
- Proof by Čencov in 1981 that Fisher metric is the *unique* metric which is invariant under so-called Markov morphisms.

Datasets and Error Distributions



- Dataset consists of observations $y_i \in \mathcal{Y} = \mathbb{R}^D$, observation conditions $x_i \in \mathcal{X} = \mathbb{R}^d$ and a specification of the uncertainty in the data points.
- For N data points, can consider the collection of all observations $\{y_i\}$ as a single point $y_{\text{data}} := (y_1, \ldots, y_N) \in \mathcal{Y}^N =: \mathcal{D}$.

Model Maps

The model map $y_{\text{model}} : \mathcal{X} \times \mathcal{M} \longrightarrow \mathcal{Y}$ must be

- sufficiently differentiable (preferably C^3) with respect to the parameters $\theta \in \mathcal{M}$,
- continuous with respect to the observation conditions $x \in \mathcal{X}$.

Can be specified

- explicitly, i.e. using a closed analytical expression
- implicitly, e.g. as the solution to a system of differential equations

$$(\mathscr{D}_x\, y_{\text{model}})(x;\theta) = f\left(x, y_{\text{model}}, \frac{\partial}{\partial x^{a_1}}\, y_{\text{model}}, \dots, \frac{\partial}{\partial x^{a_1}} \cdots \frac{\partial}{\partial x^{a_m}}\, y_{\text{model}}; \theta\right)$$

- Notations: $y_{\text{model}}(x;\theta) \equiv y(x;\theta)$.

The Likelihood Function

- Probability of measuring a given dataset if y_{model} with parameter configuration $\theta \in \mathcal{M}$ is "true".
- For independent observations, the likelihood can be factored as

$$L\Big(\mathrm{data}\,\Big|\,\theta\Big) \equiv L\Big(y_{\mathrm{data}}\,\Big|\,y_\theta(x)\Big) = \prod_{j=1}^{N} \mathbb{P}_j\Big(y_j\,\Big|\,y_{\mathrm{model}}(x_j;\theta)\Big) \qquad (4)$$

where $\mathbb{P}_j$ the error distribution associated with j-th observation.

- For correlated measurements of $\dim \mathcal{Y} = 1$ with normal error distributions

$$L\Big(\mathrm{data}\,\Big|\,\theta\Big) = \frac{1}{\sqrt{(2\pi)^N \det(\Sigma)}} \exp\left(-\frac{1}{2}\,\zeta^i(\theta)\,(\Sigma^{-1})_{ij}\,\zeta^j(\theta)\right) \qquad (5)$$

where $\zeta^a(\theta) := \Big(y_{\mathrm{data}} - h(\theta)\Big)^a$ and $h(\theta) := \Big(y(x_1;\theta), \ldots, y(x_N;\theta)\Big)$.

The Likelihood Function

- Typically it is more convenient to work with the logarithm of the likelihood $\ell := \ln(L)$.
- Gradient of log-likelihood is a covector field

$$\mathrm{d}\ell = \frac{\partial \ell}{\partial \theta^j}\,\mathrm{d}\theta^j = \frac{1}{L}\frac{\partial L}{\partial \theta^j}\,\mathrm{d}\theta^j \tag{6}$$

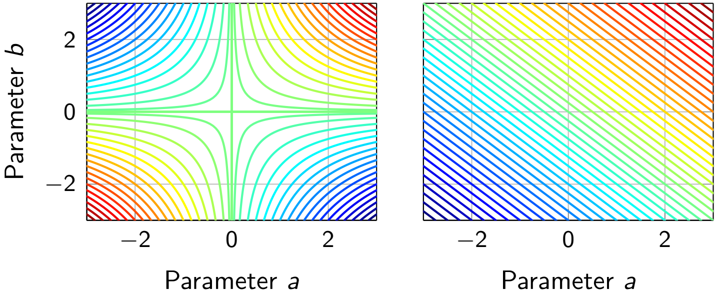
  and generally referred to as the "score".
- Maximum likelihood estimate $\theta_{\mathsf{MLE}} \in \mathcal{M}$ is defined by

$$(\mathrm{d}\ell)(\theta_{\mathsf{MLE}}) \overset{!}{=} 0 \qquad \Longleftrightarrow \qquad \left.\frac{\partial \ell}{\partial \theta^a}\right|_{\theta_{\mathsf{MLE}}} \overset{!}{=} 0. \tag{7}$$

  and $\mathrm{Hess}_\ell(\theta_{\mathsf{MLE}})$ negative-definite.

Structural Identifiability

- When is the task of finding optimal parameters well-defined?
- Counterexample: $y_{\mathrm{model}}(x; a, b) = a \cdot b \cdot x$.
- Cannot simultaneously determine both a and b from data, only the combination $m = a \cdot b$ is identifiable.
- Another example is given by $y_{\mathrm{model}}(x; a, b) = (a + b) \cdot x$ where only $m = a + b$ is identifiable.



Structural Identifiability

- Multiple desirable properties of model maps discussed e.g. in [1, 7, 8, 9].
- Existence of curves of unidentifiable configurations implies that there is a direction in $\mathcal{M}$ in which ℓ does not change.
- Most important concept of parameter identifiability is "global structural identifiability", which requires that model is injective with respect to parameters, i.e. as a map $y_{\mathrm{model}} : \mathcal{M} \longrightarrow C^0(\mathcal{X}, \mathcal{Y})$.
- Can conveniently check if model is locally injective (in a small neighbourhood around any $\theta \in \mathcal{M}$) by checking if determinant of Fisher metric is non-zero.

The Embedding Picture

- Crucial insight by Transtrum et al. in [8]: For normal likelihoods, $\mathcal{D}$ constitutes a vector space.
- View parameter manifold $\mathcal{M}$ as embedded in the "data space" $\mathcal{D} := \mathcal{Y}^N$ via the map $h : \mathcal{M} \longrightarrow \mathcal{D}$ defined by

$$h(\theta) := \Big(y_{\mathrm{model}}(x_1; \theta), \ldots, y_{\mathrm{model}}(x_N; \theta) \Big) \equiv \bigotimes_{j=1}^{N} y_{\mathrm{model}}(x_j; \theta) \in \mathcal{D}.$$

- $C : \mathcal{D} \longrightarrow \mathbb{R}_0^+$ measures distance of any point in $\mathcal{D}$ from y_{data}.

$$\begin{tikzcd}
\mathcal{M} \arrow[rr, "h = \bigotimes_{i=1}^{N} y(x_i;\theta)"] \arrow[d, "\theta"'] & & \mathcal{D} \arrow[r, "C"] \arrow[d, "w"'] & \mathbb{R}\,. \\
\theta(U) \subseteq \mathbb{R}^n \arrow[rr, "w \circ h \circ \theta^{-1}"] & & w(V) \subseteq \mathbb{R}^N \arrow[ur, "C \circ w^{-1}"'] &
\end{tikzcd}$$

Direct Consequences of Vector Space Structure on $\mathcal{D}$

- If $h : \mathcal{M} \longrightarrow \mathcal{D}$ is linear, then vector space structure of $\mathcal{D}$ is inherited along h to $\mathcal{M}$.
- Recovers Gauss–Markov theorem, which states that least squares estimator is the "best linear unbiased estimator" (BLUE).
- For $\mathcal{M}$ to be consistently embedded in $\mathcal{D}$, one must have that

$$\forall X, Y \in T\mathcal{M} : \quad g_{\mathcal{M}}(X, Y) = h^{*} g_{\mathcal{D}}(X, Y) = g_{\mathcal{D}}(h_{*}X, h_{*}Y).$$

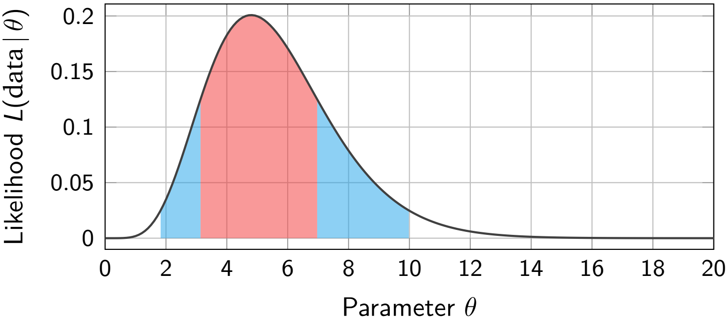
- Metric tensor on $\mathcal{D}$ given by inverse covariance between observations Σ^{-1} and is in particular constant.
- Induces distance function on $\mathcal{D}$ via

$$d_{\mathcal{D}}(\tilde{y}, y_{\text{data}}) = \sqrt{(y_{\text{data}} - \tilde{y})^{\top} \Sigma^{-1} (y_{\text{data}} - \tilde{y})} =: C(\tilde{y}). \qquad (8)$$

- Since $\mathcal{D}$ is flat, so is $\mathcal{M}$ for linear h.

Construction of Exact Confidence Regions

Confidence Intervals in 1D



- Confidence intervals should contain only the most likely parameter configurations (particularly the MLE).
- Should contain a fixed probability volume $q \in [0, 1]$.
- In 1D, one can quantify uncertainty as $\mathcal{C}_{1\sigma} = [\theta_l, \theta_u] \subseteq \mathcal{M}$.

Approximating Higher-Dimensional Confidence Regions

- For $\dim \mathcal{M} > 1$, cannot treat the uncertainties in each individual parameter independently.
- Usually, one resorts to approximations of the "true" uncertainty.
- Most popular: Estimate covariance in parameters via Cramér–Rao lower bound:

$$\Sigma_{\text{true}} \geq g^{-1}(\theta_{\text{MLE}}) \quad :\Longleftrightarrow \quad \Sigma_{\text{true}} - g^{-1}(\theta_{\text{MLE}}) \text{ positive-definite.}$$

- However, no guarantee *if or when* this lower bound is actually attained!
- For example, $a = 5 \pm 0.5$, $b = 1 \pm 0.3$ is typically not an accurate reflection of uncertainty.
- Several examples of this lower bound estimate will be shown later.

Likelihood Ratio Test and Wilks' Theorem

- Neyman–Pearson lemma (see [5]) states that likelihood ratio test is most powerful test to discriminate between simple hypotheses.
- Wilks' theorem: likelihood ratio test asymptotically distributed as

$$-2\Big(\ell(\theta) - \ell(\theta_{\mathsf{MLE}})\Big) \sim \chi^2_k \qquad (\text{as } N \longrightarrow \infty) \tag{9}$$

  with k the number of parameters in which θ and θ_{MLE} differ.
- Pearson χ^2 test statistic as second order approximation to likelihood ratio test

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y_{\mathrm{model}}(x_i;\theta)}{\sigma_i} \right)^2 . \tag{10}$$

- Often used to judge quality of a fit.

Confidence Regions via Hypothesis Tests

- If distribution of a hypothesis test is known, it can be used to define confidence regions.
- Find a threshold value below or above which points belong to confidence region.
- For likelihood-based confidence regions of level $q \in [0, 1]$ one has

$$\mathcal{C}_q := \left\{ \theta \in \mathcal{M} \;\middle|\; \ell(\theta_{\mathsf{MLE}}) - \ell(\theta) \leq \frac{1}{2} F_k^{-1}(q) \right\} \tag{11}$$

  where F_k denotes the cumulative distribution of the χ_k^2 distribution.
- The boundary of a confidence region $\mathcal{C}_q$ is then defined by

$$\partial\mathcal{C}_q = \left\{ \theta \in \mathcal{M} \;\middle|\; \ell(\theta_{\mathsf{MLE}}) - \ell(\theta) = \frac{1}{2} F_k^{-1}(q) \right\}. \tag{12}$$

Efficient Construction of Exact Confidence Regions

- Idea: Exploit the fact that confidence boundaries are level sets of a (differentiable) function f.
- Systematically construct families of vector fields which are tangential to the level sets of f.
- Then the level sets are obtained as integral manifolds (i.e. curves or surfaces) to these vector fields.
- Problem of finding confidence region is converted into (numerically) solving a system of ODEs.
- Initial condition for this system of ODEs is given by a single point which is already known to lie on the desired boundary.

Likelihood-Annihilating Vector Fields

- Given the log-likelihood function ℓ, find non-vanishing vector fields X such that

$$(\mathrm{d}\ell)(X) \equiv \mathcal{L}_X \, \ell = X^j \, \frac{\partial \ell}{\partial \theta^j} \overset{!}{=} 0 \qquad \text{everywhere.} \tag{13}$$

- One possible strategy: construct vector fields X according to

$$X^j = \alpha^j \prod_{i \neq j} \frac{\partial \ell}{\partial \theta^i} \quad \text{where} \quad \alpha^j \in \mathbb{R} : \sum_{j=1}^{\dim \mathcal{M}} \alpha^j = 0. \tag{14}$$

- Thus, the original criterion is satisfied by choosing the coefficients α^j such that $\sum_j \alpha^j = 0$ since

$$X^j \, \frac{\partial \ell}{\partial \theta^j} = \left(\sum_{j=1}^{\dim \mathcal{M}} \alpha^j \right) \underbrace{\prod_{i=1}^{\dim \mathcal{M}} \frac{\partial \ell}{\partial \theta^i}}_{=:B} = 0 \tag{15}$$

Integral Manifolds & Frobenius' Theorem

- Frobenius' thm: Every finite-dimensional Lie algebra of vector fields gives rise to a unique family of integral submanifolds (e.g. curves, surfaces) to which this Lie algebra constitutes the tangent bundle

- This family of integral submanifolds constitutes a foliation of the underlying manifold

- Set of likelihood-annihilating vector fields given by

$$\mathfrak{L} = \left\{ \left(\alpha^j \prod_{i \neq j} \frac{\partial \ell}{\partial \theta^i} \right) \frac{\partial}{\partial \theta^j} \in \Gamma(T\mathcal{M}) \;\middle|\; \sum_{j=1}^{\dim \mathcal{M}} \alpha^j = 0 \right\}. \tag{16}$$

- If one can show that $\big(\mathfrak{L}, [\,\cdot\,,\,\cdot\,]\big)$ forms a closed Lie subalgebra of $\big(\Gamma(T\mathcal{M}), [\,\cdot\,,\,\cdot\,]\big)$, then confidence boundaries $\partial \mathcal{C}_q$ foliate $\mathcal{M}$.

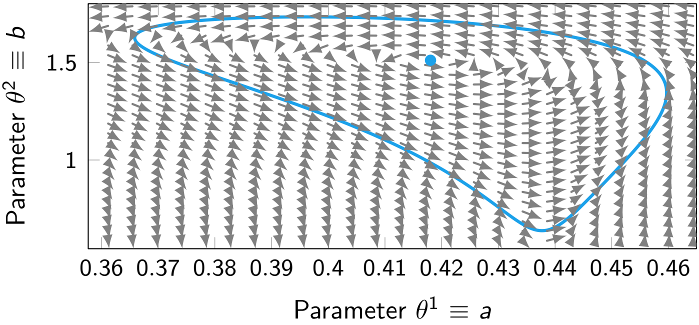
- Can be proven that this is indeed the case, provided that the model is globally structurally identifiable!

Integral Curves of Vector Fields

- For $\dim \mathcal{M} = 2$, confidence regions around MLE are topological disks and their boundaries are closed curves.

- Integral curve γ of a vector field X characterised by

$$X_{\gamma(t)} \overset{!}{=} \dot{\gamma}(t) \qquad \iff \qquad \left(X_{\gamma(t)}\right)^j \overset{!}{=} \left(\theta^j \circ \gamma\right)'(t). \tag{17}$$

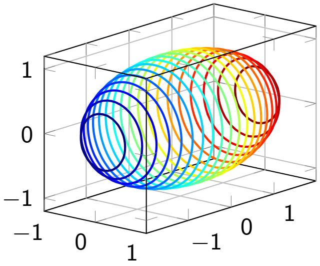


Integral Surfaces of the Likelihood

- Flows associated with likelihood-annihilating vector fields $X \in \mathfrak{L}$ convenient way of getting from any point on $\partial\mathcal{C}_q$ to any other.
- Lie algebra parametrised by coefficients $\vec{\alpha} \in \mathbb{R}^{\dim\mathcal{M}}$

$$\mathcal{H} := \left\{ \vec{\alpha} \in \mathbb{R}^{\dim\mathcal{M}} \;\middle|\; \textstyle\sum_j \alpha^j = 0 \right\}. \tag{18}$$

- $\mathcal{H}$ constitutes linear sub-vector space of $\mathbb{R}^{\dim\mathcal{M}}$, which can easily be given an orthonormal basis.



Pointwise Confidence Bands

- Also want to quantify uncertainty in prediction of model.
- Use so-called pointwise confidence bands, which are typically defined by two enveloping functions $l(x)$ and $u(x)$

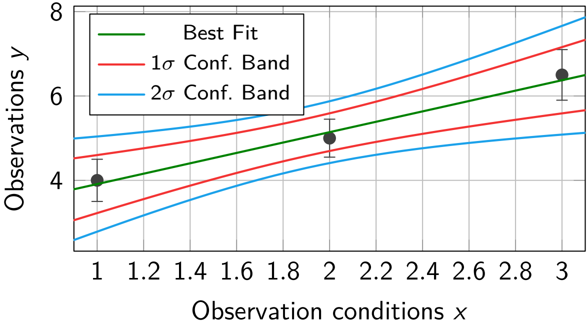
$$\forall x \in \mathcal{X} : \qquad \mathbb{P}\Big(l(x) \leq y_{\text{model}}(x; \theta_{\mathsf{MLE}}) \leq u(x)\Big) = q. \tag{19}$$

- Instead, can also define pointwise confidence bands $\mathcal{B}_q(x)$ via

$$\mathcal{B}_q(x) := y_{\text{model}}(x; \mathcal{C}_q) = \left\{ y_{\text{model}}(x; \theta) \in \mathcal{Y} \;\middle|\; \theta \in \mathcal{C}_q \right\}. \tag{20}$$

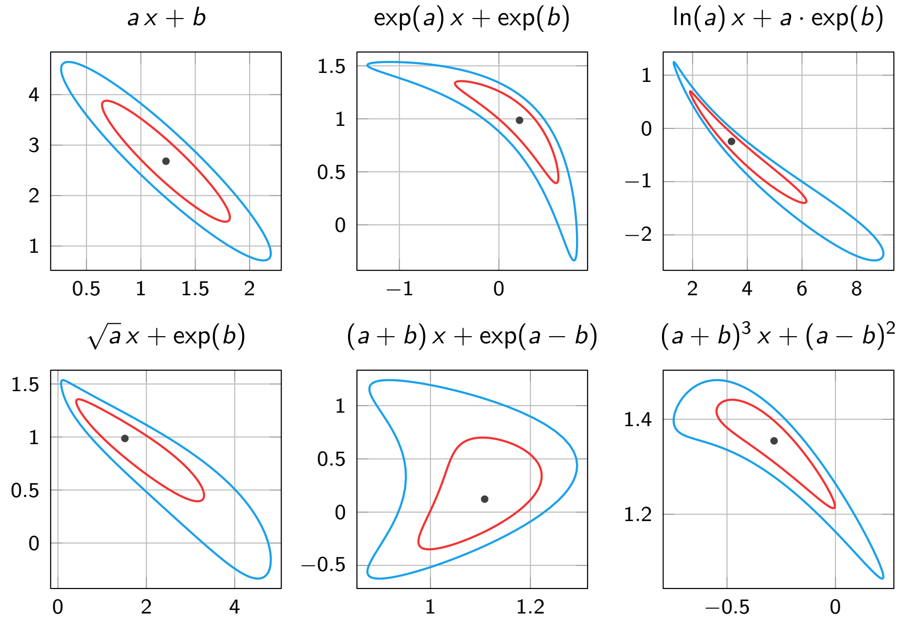
- For visualisation, one is particularly interested in boundary $(\partial\mathcal{B}_q)(x)$.
- Does $(\partial\mathcal{B}_q)(x) = \partial y_{\text{model}}(x; \mathcal{C}_q) \overset{?}{=} y_{\text{model}}(x; \partial\mathcal{C}_q)$ hold?
- Yes, guaranteed if model injective and continuous with respect to parameters! Significant reduction in computational effort.

Survey of Qualitative Effects of Reparametrisations

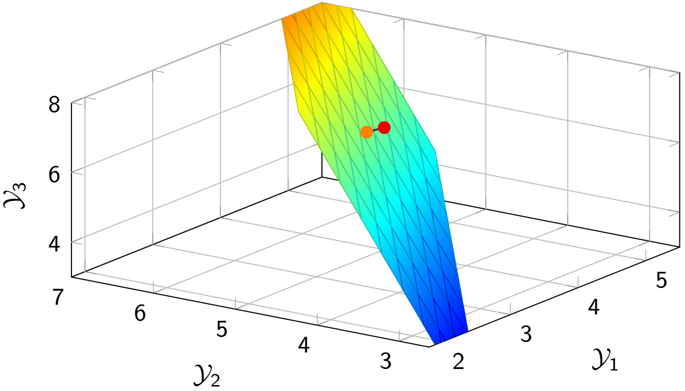


x	y	σ
1	4	0.5
2	5	0.45
3	6.5	0.6

- Confidence bands narrower in the midrange of the observation conditions of a dataset.
- Use different parametrisations of model reflecting linear relationship between x and y.
- Look at qualitative difference in shapes of confidence regions.

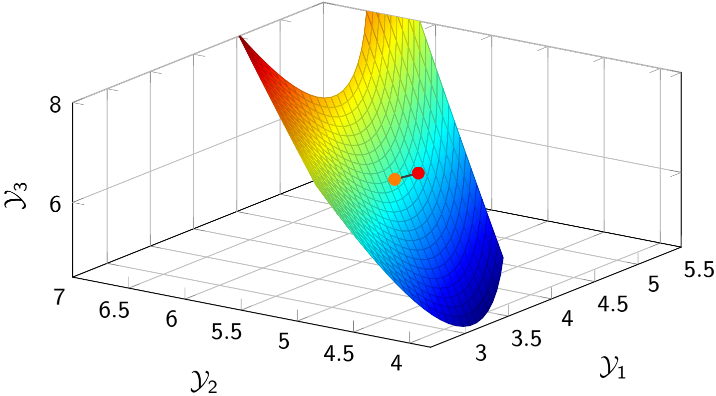

$a\,x + b$
$\exp(a)\,x + \exp(b)$
$\ln(a)\,x + a \cdot \exp(b)$
$\sqrt{a}\,x + \exp(b)$
$(a+b)\,x + \exp(a-b)$
$(a+b)^3\,x + (a-b)^2$
4
3
2
1
0.5
1
1.5
2
1.5
1
0.5
0
−1
0
1
0
−1
−2
2
4
6
8
1.5
1
0.5
0
0
2
4
1
0.5
0
−0.5
1
1.2
1.4
1.2
−0.5
0

Embedding of linearly-parametrised $\mathcal{M}$



- Uses linear parametrisation according to $y_{\text{model}}(x; a, b) = a\,x + b$.
- Red point corresponds to $y_{\text{data}} \in \mathcal{D}$, whereas orange point is $h(\theta_{\text{MLE}}) \in \mathcal{D}$.

Embedding of non-linearly parametrised $\mathcal{M}$



- Uses parametrisation $y_{\mathrm{model}}(x; a, b) = (a + b)\, x + \exp(a - b)$
- Manifold $h(\mathcal{M}) \subseteq \mathcal{D}$ is the same, coordinatisation different.

Applications of Information Geometry

Toy Model

- Linear parametrisation of quartic x–y relationship via

$$y_{\text{lin}}(x;\theta) = y_{\text{lin}}(x;a,b) = a\,x^4 + b. \tag{21}$$

- Alternatively, non-linear parametrisation according to

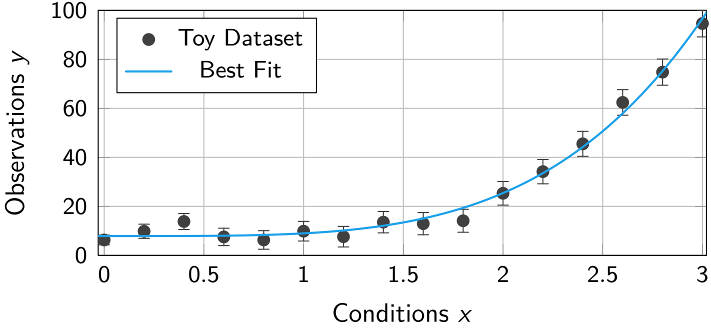
$$y_{\text{non-lin}}(x;\theta) = y_{\text{exp}}(x;a,b) = 15a^3\,x^4 + b^5. \tag{22}$$

- Reparametrisation corresponds to effective transformation on $\mathcal{M}$

$$\Phi(a,b) = (15a^3, b^5) \quad \text{and} \quad \Phi^{-1}(a,b) = \left(\sqrt[3]{a/15}, \sqrt[5]{b}\right). \tag{23}$$

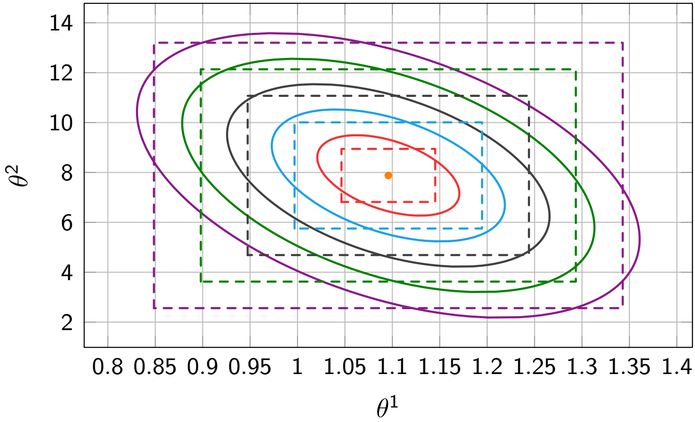
- Clearly, Φ^{-1} is no longer differentiable for $a=0$ or $b=0$.
- Study parameter manifold in both parametrisations and investigate effects of non-linear transformation.

Toy Dataset



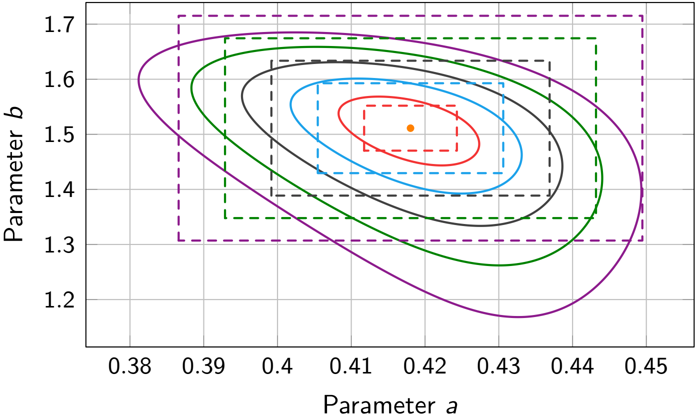
- Constant metric $g_{\mathcal{M}}$ for linear parametrisation.
- Equidistant points from y_{data} form hyperellipse in $\mathcal{D}$.
- Thus pull-back along linear h^* also results in elliptic confidence regions on $\mathcal{M}$.

Exact Confidence Regions of Linear Toy Model



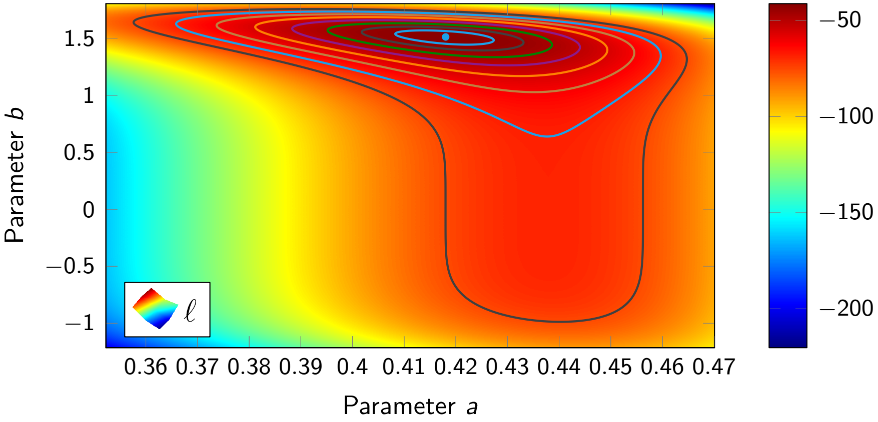
- Confidence regions from 1σ to 5σ in linear parametrisation.
- Dashed rectangles circumscribe linear estimate for covariance matrix from Cramér–Rao lower bound.

Exact Confidence Regions of Non-Linear Toy Model



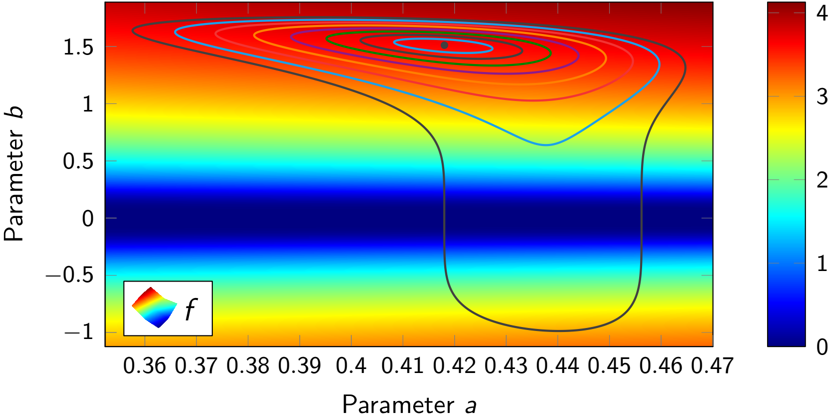
- Confidence regions from 1σ to 5σ in non-linear parametrisation.
- Can see that non-linearity introduced by chart transition Φ is relatively mild.

Log-Likelihood in Coordinates



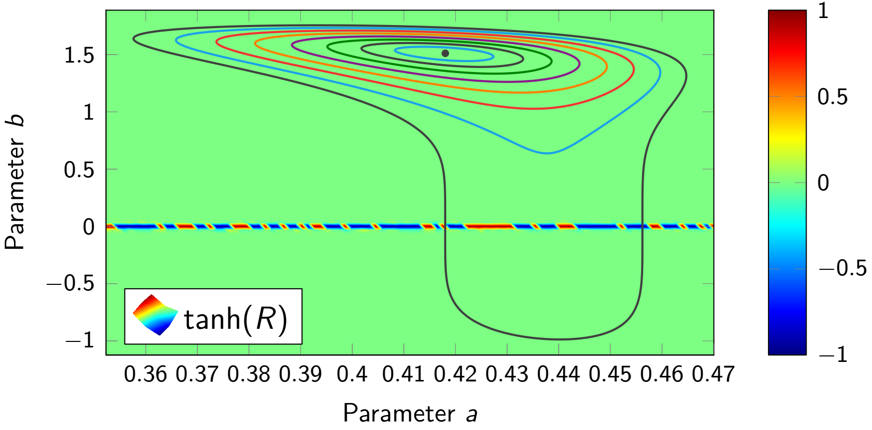
- Larger view of log-likelihood ℓ and confidence regions 1σ to 8σ.
- Not visible from log-likelihood that there is any problem at $b = 0$.

Geometric Density in Coordinates



- Rescaling according to $f = \log_{10}\left(1 + \sqrt{\det(g)}\right)$.
- Geometric density vanishes precisely on the $b = 0$ line.

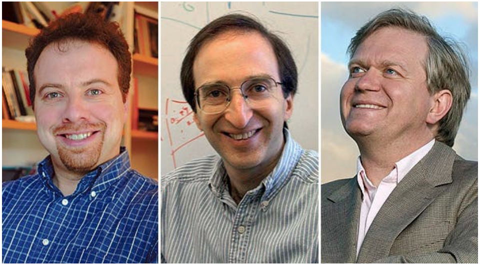
Curvature on the Parameter Manifold



- Plot of $\tanh\big(R(\theta)\big)$ reveals that $\mathcal{M}$ basically flat everywhere.
- However, one can see that $b \approx 0$ is problematic.

The Supernova Cosmology Project (SCP)

- Founded in 1988 at Berkeley National Laboratory by S. Perlmutter.
- Exploits predictable luminosity of type 1A supernova detonations to estimate their distance via their apparent brightness.
- Heads of SCP and HZT were jointly awarded the 2011 Nobel Prize in Physics for the discovery that the expansion of the Universe is accelerating.



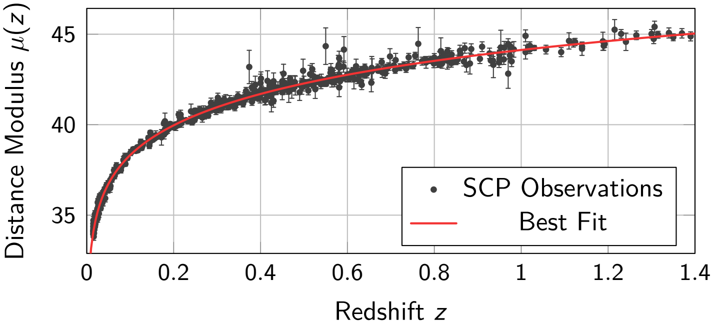
Left to right [6]:

Adam G. Riess (1/4)

Saul Perlmutter (1/2)

Brian P. Schmidt (1/4)

The SCP Dataset



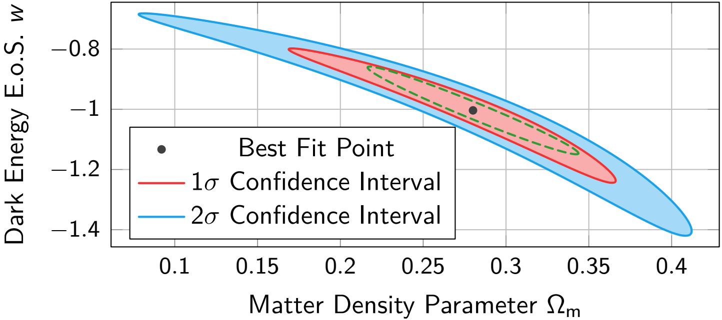
$$\mu(z;\Omega_\mathrm{m},w) = 10 + 5\log_{10}\Big[(1+z)\,d_\mathrm{H}\,\mathcal{I}(z;\Omega_\mathrm{m},w)\Big] \tag{24}$$

$$\mathcal{I}(z;\Omega_\mathrm{m},w) = \int_0^z \mathrm{d}x\ \frac{1}{\sqrt{\Omega_\mathrm{m}\,(1+x)^3 + (1-\Omega_\mathrm{m})(1+x)^{3(1+w)}}} \tag{25}$$

Physical Assumptions of the Employed Model

- Universe filled with only two cosmological fluids:
  - Matter fluid with density Ω_{m} and associated equation of state parameter $w_{\mathrm{m}} = 0$.
  - Dark energy fluid with density Ω_{Λ} and equation of state parameter w.
- Particularly, the curvature fluid Ω_K is assumed to vanish, i.e. Universe flat.
- Contribution of radiation is negligible for late cosmological times, i.e. set $\Omega_{\mathrm{rad}} = 0$.
- Thus, by Friedmann equations: $\Omega_{\mathrm{m}} + \Omega_{\Lambda} \stackrel{!}{=} 1$.
- Only need to model matter density Ω_{m} and equation of state parameter for dark energy fluid w.
- Assuming $H_0 = 70\,\frac{\mathrm{km}}{\mathrm{s \cdot Mpc}}$.

Confidence Regions for the SCP Dataset



- Non-linear dependence of model on parameters clearly visible from distorted shape of confidence regions.
- $\mathrm{vol}(\mathcal{C}_{1\sigma}) = 51.4 \pm 0.1$ and $\mathrm{vol}(\mathcal{C}_{2\sigma}) = 492 \pm 20$.

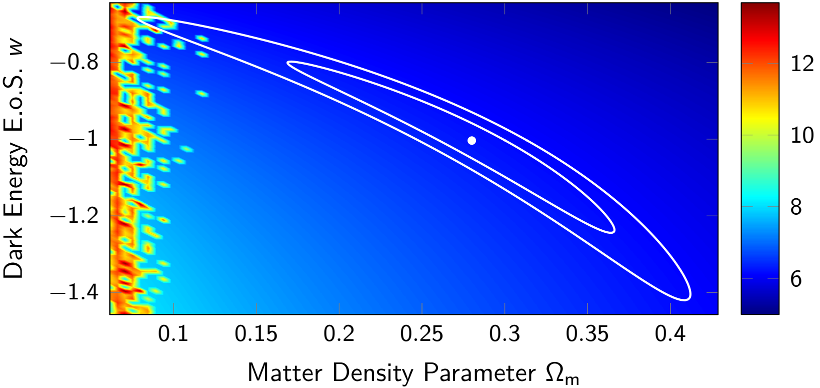
Structural Identifiability of SCP Model

- What about confidence regions of higher level?
- Closer examination of SCP model reveals that maximal open injective domain is given by

$$\mathcal{M} = \Big\{(\Omega_{\mathrm{m}}, w) \in \mathbb{R}^2 \;\Big|\; 0 < \Omega_{\mathrm{m}} < 1, w < 0\Big\}. \tag{26}$$

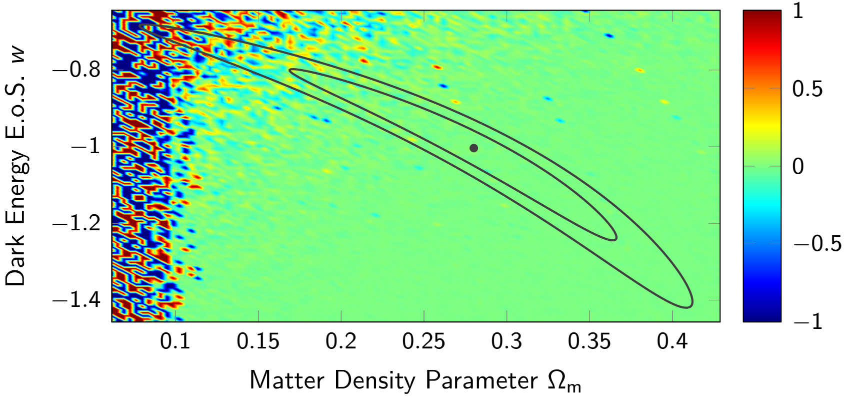
- At approximately 2.734σ, the boundary of the confidence region touches the $\Omega_{\mathrm{m}} = 0$ line.
- Conceptual upper limit to confidence regions which can be displayed in this chart.

Geometric Density in Coordinates



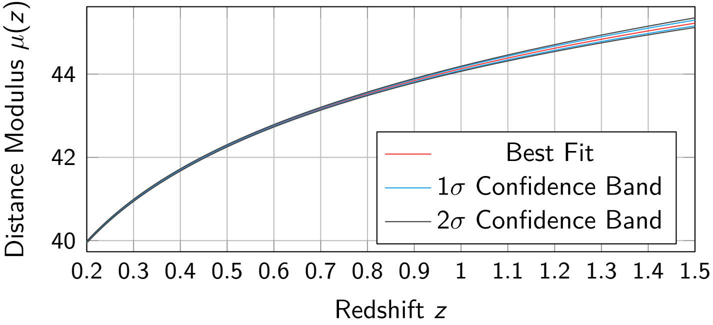
- Rescaled geometric density factor according to $\ln\left(\sqrt{\det g}\right)$.
- Strong fluctuations in geometric density factor for small Ω_m.
- Due to rapid growth in components of Fisher metric.

Curvature on the Parameter Manifold



- Rescaled curvature as $f = \tanh(8 \tanh R)$.
- Similar to geometric density, fluctuating curvature near $\Omega_m = 0$.
- Indication of chart boundary as with non-linear toy model?

Confidence Bands for SCP

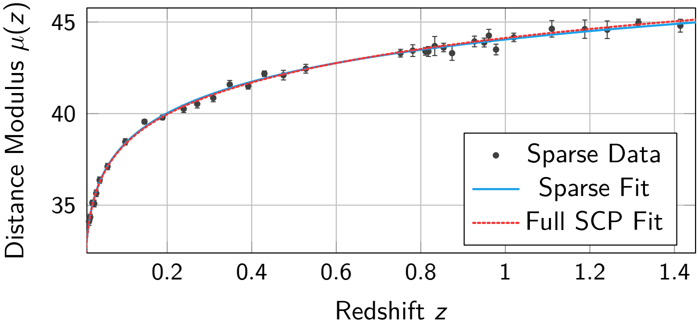


- Increasing width of confidence band for larger redshifts z.
- New observations at high z will constrain uncertainty in prediction more than observations at lower z.

Sparse Excerpt of SCP Dataset

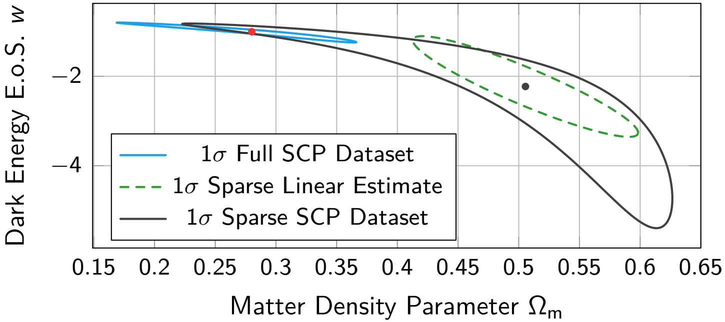
- To verify consistency of integral manifold method, study sparse excerpt of SCP dataset.
- Clearly, MLE obtained from smaller dataset will be different.
- Does the MLE of full dataset still lie within confidence region associated with excerpt?
- What at what confidence level? 1σ? 2σ?
- What about linear approximation of uncertainty for sparse MLE?

Sparsified SCP Dataset



- Semi-randomly chosen subset containing $35/580 \approx 6\,\%$ of the SCP dataset.
- Although the prediction looks almost indistinguishable, one finds $\theta_{\text{MLE,Sparse}} \approx (0.51, -2.23)$ compared with $\theta_{\text{MLE,Full}} \approx (0.28, -1.00)$.

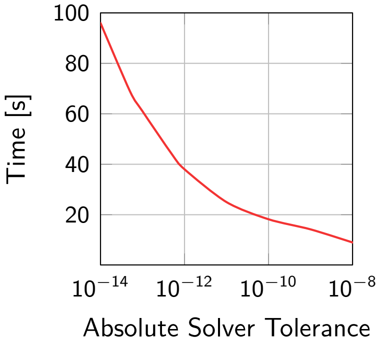
Confidence Regions of Sparse SCP Dataset



- For CRLB, the "true" MLE apparently far outside 1σ region, whereas exact region shows that this is still within 1σ.
- Therefore, CRLB grossly misrepresents true uncertainty in parameters both in shape and magnitude. (Not even a true lower bound!)

Performance for the SCP dataset (on my tablet)

Solver tol.	#Function Evals	Time
10^{-8}	740	9.0 s
10^{-9}	1100	14.2 s
10^{-10}	1500	18.2 s
10^{-11}	2100	25 s
10^{-12}	3200	38 s
$5 \cdot 10^{-13}$	3600	44 s
10^{-13}	5000	61 s
$5 \cdot 10^{-14}$	5700	69 s
10^{-14}	8000	96 s



- Single core performance of integral curve scheme for 1σ boundary of the SCP dataset.
- Score evaluation takes 8.8 ms, log-likelihood takes 0.88 ms.

Comparison of Complexities

- Naïve method of constructing exact confidence regions: evaluate likelihood ratio test on a grid of parameter configurations and use interpolation to estimate location of boundary.
- Scales according to $O(H^{\dim \mathcal{M}})$ where H denotes the grid density.
- Majority of likelihood evaluations "wasted" far away from location of boundary.
- Integral manifold method requires evaluation of gradient of likelihood (which has $\dim \mathcal{M}$ components) at every step.
- Scales according to $O(\dim \mathcal{M} \cdot H^{\dim \mathcal{M}-1})$ where H denotes the grid density.
- Significantly more efficient use of likelihood evaluations.

Summary

- Parameter space $\mathcal{M}$ should be viewed as a manifold instead of as a vector space.
- Valid domain of a parametrised model can be conveniently investigated using manifold invariants.
- Especially for non-linearly parametrised models, Cramér-Rao lower bound yields poor approximation of both shape and magnitude of true uncertainty.
- For "well-defined" models, exact confidence boundaries and confidence bands both exist and can be computed efficiently.

Further Results

- Geometries induced on $\mathcal{D}$ by non-normal error distributions associated with observations.
- For example, metric induced by pseudo-Poisson errors results in non-vanishing Christoffel symbols on $\mathcal{D}$ (but still $R = 0$).
- Cauchy error distributions result in “halved information content” of each observation.
- Square of geodesic distance on $\mathcal{M}$ can be used as a reliable approximation to likelihood ratio test in cases of normal error distributions.

Outlook and Strict Subset of Open Questions

- Implement non-normal (e.g. asymmetric) error distributions and investigate associated confidence regions.
- Extend formalism to include uncertainty in conditions $x \in \mathcal{X}$ (see e.g. [3, 4]).
- If $\mathcal{M}$ flat, there should exist a smooth coordinate transformation in which makes confidence regions elliptic (see also [2]).
- Study Lie group associated with algebra of likelihood-annihilating vector fields (e.g. its Killing form).
- Simultaneous confidence bands in geometric picture?
- Bayesian analogue of geometric Fisher formalism?
- Finsler Geometry instead of Riemannian Geometry: can it provide a better approximation to Kullback–Leibler divergence?

Questions?

✉ arutjunjan.r@gmail.com

/RafaelArutjunjan/Master-Thesis

/RafaelArutjunjan/InformationGeometry.jl

References I

[1] ARUTJUNJAN, R. : *On the Geometric Foundation of Parameter Inference*, Friedrich-Alexander University Erlangen-Nürnberg, Master's thesis, August 2020. https://github.com/RafaelArutjunjan/Master-Thesis

[2] GIESEL, E. S.: *Investigation of Non-Gaussian Likelihoods in the Framework of Information Geometry*, Heidelberg University, Master's Thesis

[3] HEAVENS, A. F. ; SEIKEL, M. ; NORD, B. D. ; AICH, M. ; BOUFFANAIS, Y. ; BASSETT, B. A. ; HOBSON, M. P.: Generalised Fisher Matrices. (2014). https://arxiv.org/pdf/1404.2854v2.pdf

[4] HEAVENS, A. : Generalisations of Fisher Matrices. In: *Entropy* 18 (2016), Jun, Nr. 6, 236. http://dx.doi.org/10.3390/e18060236. – DOI 10.3390/e18060236. – ISSN 1099–4300

[5] NEYMAN, J. ; PEARSON, E. S. ; PEARSON, K. : IX. On the problem of the most efficient tests of statistical hypotheses. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231 (1933), Nr. 694-706, 289-337. http://dx.doi.org/10.1098/rsta.1933.0009. – DOI 10.1098/rsta.1933.0009

References II

[6] Overbye, D. : Studies of Universe's Expansion Win Physics Nobel. (2011), October. https://www.nytimes.com/2011/10/05/science/space/05nobel.html

[7] Transtrum, M. K. ; Machta, B. ; Brown, K. ; Daniels, B. C. ; Myers, C. R. ; Sethna, J. P.: Sloppiness and Emergent Theories in Physics, Biology, and Beyond. (2015). https://arxiv.org/pdf/1501.07668.pdf

[8] Transtrum, M. K. ; Machta, B. B. ; Sethna, J. P.: Geometry of nonlinear least squares with applications to sloppy models and optimization. 83 (2011), March, Nr. 3, S. 036701. http://dx.doi.org/10.1103/PhysRevE.83.036701. – DOI 10.1103/PhysRevE.83.036701

[9] White, A. ; Tolman, M. ; Thames, H. D. ; Withers, H. R. ; Mason, K. A. ; Transtrum, M. K.: The Limitations of Model-Based Experimental Design and Parameter Estimation in Sloppy Systems. In: *PLOS Computational Biology* 12 (2016), 12, Nr. 12, 1-26. http://dx.doi.org/10.1371/journal.pcbi.1005227. – DOI 10.1371/journal.pcbi.1005227