On the Geometric Foundation of Parameter Inference

Master's Thesis in Physics

Presented by

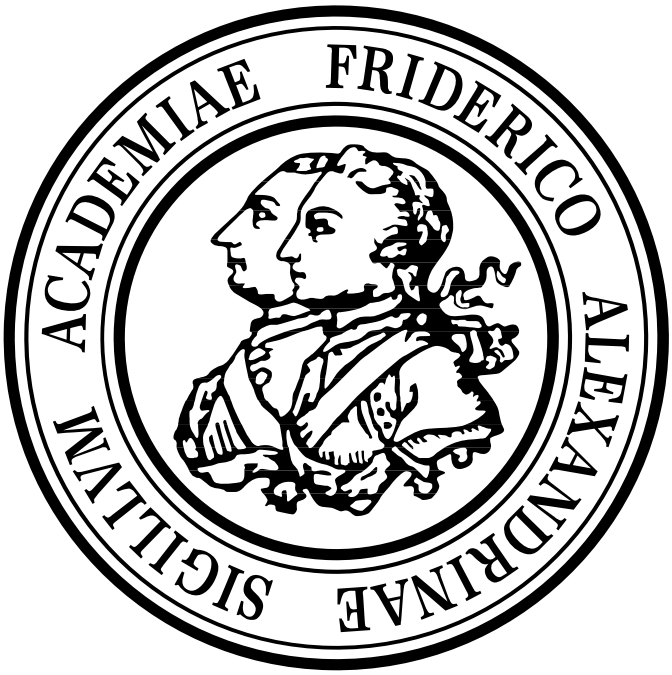
Rafael Arutjunjan

on the 24th of August 2020

Last Correction: October 22, 2020

Chair of Theoretical Physics I

Friedrich-Alexander University Erlangen-Nürnberg



Supervisor: Prof. Dr. Klaus Mecke

Co-Supervisor: Prof. Dr. Björn Malte Schäfer

To my mother

Contents

Abstract

An overview of methods from differential geometry, probability theory as well as their intersection in the subject of information geometry is given. Subsequently, an extension to the established geometric framework of parameter inference is proposed and its implications for the geometry on the data space explored in various different settings.

A generalisation of confidence regions to parameter manifolds is presented and an efficient exact method for their construction developed. A proof which highlights the necessary conditions on the likelihood function such that the confidence boundaries foliate the parameter manifold is given. Lastly, it is shown how confidence regions can be propagated to establish so-called confidence bands, which in turn quantify the uncertainty in the predictions of a model.

Lastly, to demonstrate their real-world capabilities, the developed geometric methods are applied to cosmological observations of type 1A supernovæ.

Zusammenfassung

Zu Beginn wird eine Übersicht der wichtigsten Methoden der Differenzialgeometrie, Wahrscheinlichkeitstheorie und ihrer Kombination durch das mathematische Gebiet der Informationsgeometrie gegeben. Eine Erweiterung des bekannten informationsgemetrischen Formalismus wird vorgeschlagen und ihre Konsequenzen im Bezug auf die Geometrie des Datenraumes in mehreren Beispielen untersucht.

Das Konzept der Konfidenzintervalle wird auf Parametermannigfaltigkeiten verallgemeinert und eine effiziente Methode für die Berechnung exakter Konfidenzregionen entwickelt. Im Rahmen eines Beweises werden hinreichende Bedingungen an die sogenannte Likelihood-Funktion ersichtlich, unter welchen die Grenzflächen von Konfidenzregionen eine Blätterung der Parametermannigfaltigkeit darstellen. Darüber hinaus wird aufgezeigt, wie gegebene Konfidenzregionen dazu verwendet werden können, die Unsicherheit in den Vorhersagen eines Modells zu quantifizieren.

Zuletzt werden die entwickelten Methoden zur Untersuchung kosmologischer Messungen von Supernovæ des Typs 1A verwendet.

1 Introduction

Essentially, the goal of science is to provide a quantitative understanding of natural phenomena by constructing models whose predictions can be tested and falsified by experiment. If a model repeatedly withstands falsification by experimental tests and even correctly predicts previously unobserved phenomena, it is hoped that the model has captured at last a part of the underlying mechanisms that govern these phenomena.

Therefore, the primary purpose of a model or hypothesis is to provide a concise abstraction which distils large numbers of experimental observations into an explicit functional relationship that describes the observations as best as possible. In this sense, the common phrase "a picture is worth a thousand words" can be translated as "a model is worth a thousand datasets" in the context of scientific data analysis.

The aim of information geometry is to apply the powerful methods provided by the mathematical discipline of differential geometry to parameter inference and statistical analyses in general. That is, statistical problems are rephrased in such a way that they can be given a geometric interpretation. The origins of this idea can be traced back to 1945 when C. Rao first introduced the seminal idea of establishing a Riemannian metric on a space of probability distributions (see [5]). With this, he was able to link the previously non-overlapping mathematical subjects of differential geometry and probability theory.

While the problems of finding optimal parameters, performing systematic model reduction and designing optimal experiments have been discussed by numerous publications so far, the topics of uncertainty in the parameters and confidence regions have generally not been addressed in the available information-geometric literature.

In particular, there appears to be no generally agreed upon way of specifying the exact uncertainty in the optimal parameter configuration. Instead, most researchers seem to rely on linear approximations of the parameter uncertainty via the Cramér–Rao lower bound. Hence, the presented work aims to remedy this by presenting a definition of confidence regions which is valid irrespective of the chosen parametrisation and by providing an efficient method with which they can be constructed in practice.

Furthermore, an overwhelming majority of the available literature only discusses systems where the uncertainties and errors associated with the observations follow a normal distribution. Usually, the central limit theorem is quoted as a justification for restricting attention to the normal case. However, the assumption that uncertainties associated with the observations perfectly follow a normal distribution can often yield an inaccurate approximation, especially in the case of small datasets. Consequently, this thesis also explores a systematic way of treating a wider class of error distributions.

2 Mathematical Preliminaries

This section aims to provide a reasonably comprehensive summary of central concepts and core definitions of differential geometry, probability theory and information geometry in so far as they are relevant to the discussions and interpretation of results later on. While this section is mainly provided for sake of completeness and the reader's convenience, it has the added benefit of clarifying and showcasing the notation that will be used throughout this thesis. Above all, there is no pretence whatsoever as to the originality of the notions which are discussed in this section.

Rigorous proofs are omitted in favour of short arguments almost everywhere in this exposition wherefore the reader is referred to the standard literature on the subjects for a more detailed treatment: for topology and differential geometry, see e.g. [12, 20, 22, 29, 33, 42, 46, 47, 65–67, 82], for statistics and probability theory see e.g. [8, 14, 17, 26, 34, 38, 39, 45, 68] and for information geometry see e.g. [3, 18, 40, 49, 52, 64, 71, 78].

For readers who are already familiar with the core concepts of differential geometry and probability theory, it is recommended to proceed to section 3 and only revisit the basic definitions presented in this exposition as needed.

2.1 Topological Spaces

This section is mainly based on [22, 46, 65, 67].

The starting point of any serious theory, mathematical or otherwise, is a set of axioms—a comprehensive list of fundamental assumptions that are accepted as true, forming the building blocks from which all other properties and arguments of the theory are derived. For modern mathematics, these underlying axioms are the so-called Zermelo–Fraenkel axioms of set theory. On top of these, the so-called axiom of choice is usually also included. Since an in-depth discussion on Zermelo–Fraenkel set theory and the implications of the axiom of choice are outside the scope of this thesis, the reader is referred to specialised literature on the subject, such as [35]. Throughout this thesis, it will be assumed that the developed mathematical formalism rests firmly on top of the set-theoretic foundation provided by the ZFC axioms (i.e. with the axiom of choice explicitly included).

Given a set of points $\mathcal{M}$, the weakest known structure it can be equipped with to provide its elements with a notion of adjacency or neighbourhood is a topology. The choice of any particular topology also determines a variety of other properties of $\mathcal{M}$, relating to continuity, connectedness, compactness, metrisability, separability and so on.

A topology $\mathcal{O} \subseteq \mathscr{P}(\mathcal{M})$ on a set $\mathcal{M}$ is itself a set of sets, subject to the conditions

(T1) $\varnothing \in \mathcal{O},\ \mathcal{M} \in \mathcal{O}$

(T2) $\forall j \in J : U_j \in \mathcal{O} : J \text{ finite} \implies \bigcap_{j\in J} U_j \in \mathcal{O}$

(T3) $\forall i \in I : U_i \in \mathcal{O} \implies \bigcup_{i\in I} U_i \in \mathcal{O}$

where $\varnothing$ and $\mathscr{P}(\mathcal{M})$ respectively denote the empty set and the power set[1] of $\mathcal{M}$, and both J

[1] The powerset $\mathscr{P}(\mathcal{M})$ of a set $\mathcal{M}$ is the collection of all subsets of $\mathcal{M}$, i.e. $\mathscr{P}(\mathcal{M}) = \{X \subseteq \mathcal{M}\}$.

and I are index sets. In words, the above requirements mean that a topology is constructed from subsets of the underlying set $\mathcal{M}$ and that finite intersections and arbitrary unions of elements in the topology must also be contained in the topology.[2]

Such a tuple $(\mathcal{M}, \mathcal{O})$ is then called a topological space. Furthermore, the elements of the topology $\mathcal{O}$ are called the open sets of $\mathcal{M}$. Also, a set $A \subseteq \mathcal{M}$ is said to be closed if its complement $A^c := \mathcal{M} \setminus A = \{x \in \mathcal{M} \mid x \notin A\}$ is open.[3]

To illustrate the requirements for a topology with a simple example, consider the finite set $\mathcal{M} = \{1, 2, 3\}$. A topology on $\mathcal{M}$ is given, for example by $\mathcal{O} = \Big\{\{1\}, \{1, 2, 3\}\Big\}$ or $\mathcal{O} = \Big\{\{2\}, \{1, 2, 3\}\Big\}$.

However, the set $\mathcal{O} = \Big\{\{1\}, \{2\}, \{1, 2, 3\}\Big\}$ does not constitute a topology on $\mathcal{M}$ since property (iii) is violated:

$$\{1\} \cup \{2\} = \{1, 2\} \notin \Big\{\{1\}, \{2\}, \{1, 2, 3\}\Big\}. \tag{2.1}$$

Also, the candidate $\mathcal{O} = \Big\{\{1, 2\}, \{2, 3\}, \{1, 2, 3\}\Big\}$ fails to satisfy (ii) since

$$\{1, 2\} \cap \{2, 3\} = \{2\} \notin \Big\{\{1, 2\}, \{2, 3\}, \{1, 2, 3\}\Big\}. \tag{2.2}$$

The empty set is generally suppressed in the notation since it is already guaranteed to be contained by all sets according to the ZFC axioms.

The two most extreme topologies that a set $\mathcal{M}$ can be equipped with are the chaotic and discrete topologies

$$\mathcal{O}_{\text{chaotic}} := \{\varnothing, \mathcal{M}\} \qquad \text{and} \qquad \mathcal{O}_{\text{discrete}} := \mathcal{P}(\mathcal{M}). \tag{2.3}$$

Neither of these topologies result in topological spaces which are useful in practical applications, however, they often serve as helpful edge cases when trying to understand and prove topological statements.

One can gather from table 1 that the freedom of choice in establishing inequivalent topologies on a set grows very rapidly with the size of the underlying set. Therefore, it would be advantageous if one were able to compare different topologies: for two topological spaces $(\mathcal{M}, \mathcal{O}_1)$ and $(\mathcal{M}, \mathcal{O}_2)$ which share the same underlying set $\mathcal{M}$, one says that $\mathcal{O}_2$ is finer than $\mathcal{O}_1$ or equivalently that $\mathcal{O}_1$ is coarser than $\mathcal{O}_2$ if $\mathcal{O}_1 \subseteq \mathcal{O}_2$. Note that neither topology need contain the other and thus not all topologies can be ordered this way. Clearly, the coarsest topology possible on any set is the chaotic topology, whereas the finest possible topology is the discrete topology.

While it is possible to provide topologies for finite spaces explicitly by writing down a complete list of its elements, one must usually resort to indirect specifications of topologies on infinite sets. An important example of this is the so-called standard topology on $\mathbb{R}^n$. First, one defines open balls

[2] The property (T1) is actually already implied by choosing $I = \varnothing = J$ in (T2) and (T3). However, since it can be easy to forget to check this case explicitly, it is usually included as a separate condition.

[3] Since a set can be open, closed, both or neither, this terminology should not be taken too literally. For example, the empty set $\varnothing$ and the whole set $\mathcal{M}$ are both open and closed with respect to any topology due to (T1). Also, half-open intervals $[a, b) \subseteq \mathbb{R}$ are neither open nor closed with respect to the standard topology on $\mathbb{R}$.

$n = \lvert\mathcal{M}\rvert$	Distinct Topologies	Inequivalent Topologies
1	1	1
2	4	3
3	29	9
4	355	33
5	6942	139
6	209 527	718
7	9 535 241	4535

Table 1: This table lists the number of distinct admissible topologies as well as the number of inequivalent topologies on a finite set $\mathcal{M}$ as a function of its cardinality. Here, inequivalent refers to the fact that the topologies result in mutually non-homeomorphic spaces. The numerical values are excerpted from the Online Encyclopedia of Integer Sequences [73] and correspond to the labels A000798 and A001930 respectively.

$B_r(p)$ of radius r around a point $p \in \mathbb{R}^d$ via

$$B_r(p) \coloneqq \left\{ (q_1, ..., q_d) \in \mathbb{R}^d \;\middle|\; \sum_{i=1}^{d} (q_i - p_i)^2 < r^2 \right\}. \tag{2.4}$$

Then the standard topology is defined as

$$\mathcal{O}_{\text{std}} \coloneqq \left\{ U \subseteq \mathbb{R}^d \;\middle|\; \forall p \in U : \exists r > 0 : B_r(p) \subseteq U \right\}. \tag{2.5}$$

That is, a set $U \subseteq \mathcal{M}$ is defined to be open in the standard topology if around every point $p \in U$ there exists a non-empty open ball which is still entirely contained in U.[4] An illustration of this is given in figure 1. It is usually implied that the standard topology on $\mathbb{R}^d$ is chosen by default, unless the space $\mathbb{R}^d$ is additionally equipped with some other structure, from which the topology can be inherited.

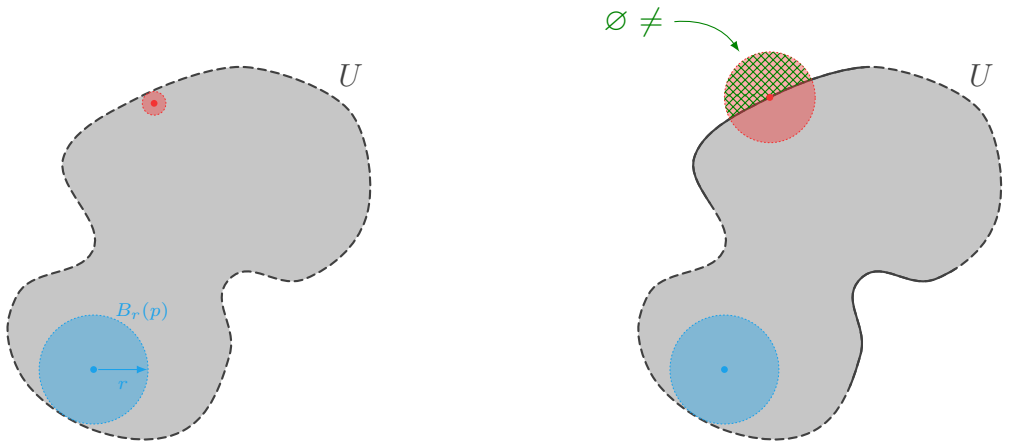
In the field of topology, the preimage of a map f is commonly denoted by f^{-1} where it is understood that this is not supposed to imply that the map f is invertible. However, since this notation can lead to confusion and misunderstandings when mixing topology with other areas of mathematics, the more explicit notation of preim_f will be used to indicate the preimage here.

Suppose $(\mathcal{M}, \mathcal{O}_\mathcal{M})$ and $(\mathcal{N}, \mathcal{O}_\mathcal{N})$ are topological spaces. Then a map $f : \mathcal{M} \longrightarrow \mathcal{N}$ between them is defined to be continuous if

$$\forall U \in \mathcal{O}_\mathcal{N} : \qquad \left\{ x \in \mathcal{M} \;\middle|\; f(x) \in U \right\} =: \text{preim}_f(U) \in \mathcal{O}_\mathcal{M}, \tag{2.6}$$

which means that all preimages of open sets under the map f must again be open sets. That is,

[4] Showing that $\mathcal{O}_{\text{std}}$ indeed constitutes a valid topology is a standard proof which can be found in the listed references.



(a) Example of an open set in $(\mathbb{R}^2, \mathcal{O}_{\text{std}})$.

(b) Example of a set in $(\mathbb{R}^2, \mathcal{O}_{\text{std}})$ that is neither open nor closed.

Figure 1: Visualisation of open sets of $\mathbb{R}^2$ with respect to the standard topology. Dashed lines indicate parts of the boundary that do not belong to the enclosed set whereas solid lines conversely represent elements on the boundary that do belong to the enclosed set. One can see that for points on the boundary, there exists no $r > 0$ such that a ball B_r around them still lies entirely in the set. Although the concept of distance has been left undefined so far, the intuition given by the visualisations is correct nonetheless.

continuous maps between topological spaces preserve the open set structure, i.e. the topologies.[5]

A map $f : \mathcal{M} \longrightarrow \mathcal{N}$ between two topological spaces $(\mathcal{M}, \mathcal{O}_\mathcal{M})$ and $(\mathcal{N}, \mathcal{O}_\mathcal{N})$ is said to be a homeomorphism if it is invertible and continuous in both directions. Since they preserve the topological structure both ways, homeomorphisms are isomorphisms between topological spaces. As with any other structure, spaces which are isomorphic are said to be equivalent, i.e. topological spaces between which a homeomorphism exists are considered topologically equivalent.

2.1.1 Inheriting Topologies

Since it is preferable to be economical in the introduction of new structure, one would ideally like to inherit structures from one space to another whenever possible. Frequent cases of this include inheriting a topology from a topological space $(\mathcal{M}, \mathcal{O})$ to

1. a subspace $\mathcal{N} \subseteq \mathcal{M}$,
2. an initial space $\mathcal{N}$ connected to $\mathcal{M}$ via $f : \mathcal{N} \longrightarrow \mathcal{M}$,
3. a final space $\mathcal{N}$ connected to $\mathcal{M}$ via $g : \mathcal{M} \longrightarrow \mathcal{N}$,
4. a product or quotient space,

and so on.

For a subset $\mathcal{N} \subseteq \mathcal{M}$ of an ambient topological space $(\mathcal{M}, \mathcal{O})$, the subset topology $\mathcal{O}|_\mathcal{N}$ is

[5] It can be shown that this definition of continuity corresponds exactly to the ε–δ definition from standard analysis for functions $f : \mathbb{R} \longrightarrow \mathbb{R}$ if the standard topology is chosen on both the domain and the target space.

generated by

$$\mathcal{O}|_{\mathcal{N}} \coloneqq \{\mathcal{N} \cap U \mid U \in \mathcal{O}\}. \tag{2.7}$$

One can check that this definition satisfies the requirements for a topology. An alternative notation for the subset topology $\mathcal{O}|_{\mathcal{N}}$ that is often found in literature is $\mathcal{N} \cap \mathcal{O}$. Since properties such as openness can be fulfilled on a subset topology but not with respect to the topology of the ambient space and also the other way around, it is considered best practice to always state explicitly with respect to which topology a statement is made.

On a Cartesian product $\mathcal{M} \times \mathcal{N} = \left\{(m,n) \;\middle|\; m \in \mathcal{M}, n \in \mathcal{N}\right\}$ of two topological spaces $(\mathcal{M}, \mathcal{O}_{\mathcal{M}})$ and $(\mathcal{N}, \mathcal{O}_{\mathcal{N}})$ one can construct the product topology $\mathcal{O}_{\mathcal{M}\times\mathcal{N}}$ by

$$\mathcal{O}_{\mathcal{M}\times\mathcal{N}} \coloneqq \left\{U \subseteq \mathcal{M} \times \mathcal{N} \;\middle|\; \forall p \in U : \exists S \in \mathcal{O}_{\mathcal{M}} : \exists T \in \mathcal{O}_{\mathcal{N}} : p \in S \times T \subseteq U\right\}. \tag{2.8}$$

It can be shown that this again constitutes a valid topology and moreover that the product topology $\mathcal{O}_{\mathbb{R}\times\ldots\times\mathbb{R}}$ on $\mathbb{R}^d$ (with each $\mathcal{O}_{\mathbb{R}}$ the standard topology on $\mathbb{R}$) is equivalent to the standard topology $\mathcal{O}_{\mathbb{R}^d}$ on $\mathbb{R}^d$.

The interior $\mathrm{Int}(X)$, the closure $\overline{X}$ and the boundary ∂X of a set $X \subseteq (\mathcal{M}, \mathcal{O})$ are defined, respectively, by

$$\mathrm{Int}(X) \coloneqq \bigcup_{\substack{O \subseteq X \\ O \text{ open}}} O, \qquad \overline{X} \coloneqq \bigcap_{\substack{X \subseteq A \\ A \text{ closed}}} A, \qquad \partial X \coloneqq \overline{X} \setminus \mathrm{Int}(X). \tag{2.9}$$

From these definitions, it is not hard to see that the interior $\mathrm{Int}(X)$ is the largest possible open subset contained in X, whereas the closure $\overline{X}$ is the smallest possible closed superset which contains X.

A topological space $(\mathcal{M}, \mathcal{O})$ is said to be Hausdorff if around any pair of distinct points, there exist mutually non-intersecting open neighbourhoods. Formally,

$$\forall p, q \in \mathcal{M} : p \neq q : \exists U_p, U_q \in \mathcal{O}_{\mathcal{M}} : p \in U_p : q \in U_q : \quad U_p \cap U_q = \varnothing. \tag{2.10}$$

This Hausdorff property is only one of several separability properties that a given topological space can have. However, it may be one of the most significant since it implies many other desirable properties for a topological space. For a start, in spaces which are not Hausdorff, sequences can converge to more than one point—even all points—simultaneously.

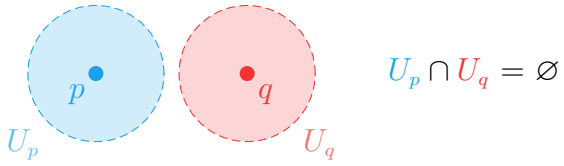


Figure 2: Illustration of the Hausdorff separation property between two points in a topological space.

Given any collection of subsets $S \subseteq \mathcal{P}(\mathcal{M})$, one can generate a topology on $\mathcal{M}$ which contains S via the construction

$$\langle S \rangle \coloneqq \bigcap_{\substack{S \subseteq \mathcal{O} \\ \mathcal{O} \text{ a topology on } \mathcal{M}}} \mathcal{O}. \tag{2.11}$$

That is, $\langle S \rangle$ denotes the coarsest possible topology on $\mathcal{M}$ which contains all elements of S. For a family $(f_i)_{i \in I}$ of maps $f_i : X \longrightarrow Y_i$, the initial topology $\mathcal{O}_{\text{initial}}$ on X induced from the topologies on the spaces Y_i is constructed via

$$\mathcal{O}_{\text{initial}} \coloneqq \left\langle \bigcup_{i \in I} \left\{ \text{preim}_{f_i}(U) \subseteq X \;\middle|\; U \in \mathcal{O}_i \right\} \right\rangle \tag{2.12}$$

which means that all maps $(f_i)_{i \in I}$ are made continuous. Furthermore, the initial topology is the coarsest possible topology under which all $f_i : X \longrightarrow Y_i$ are continuous.[6]

The final topology $\mathcal{O}_{\text{final}}$ on Y induced by a family $(g_i)_{i \in I}$ of maps $g_i : X_i \longrightarrow Y$, whose domains are topological spaces $(X_i, \mathcal{O}_i)$, is given by

$$\mathcal{O}_{\text{final}} \coloneqq \left\{ U \subseteq Y \;\middle|\; \forall i \in I : \text{preim}_{g_i}(U) \in \mathcal{O}_i \right\}. \tag{2.13}$$

The final topology is the finest possible topology on Y such that all maps $g_i : X_i \longrightarrow Y$ are rendered continuous.

2.1.2 Connectedness and Compactness

A topological space $(\mathcal{M}, \mathcal{O})$ is said to be connected if there exists no decomposition into two disjoint non-empty open sets, i.e. if for any $U_1, U_2 \in \mathcal{O}$ with $\mathcal{M} = U_1 \cup U_2$ it is true that

$$U_1 \cap U_2 = \varnothing \qquad \Longrightarrow \qquad U_1 = \varnothing \quad \text{or} \quad U_2 = \varnothing. \tag{2.14}$$

Importantly, images of connected sets under continuous maps are also connected.

An open cover C of a set $U \subseteq \mathcal{M}$ is a collection of open sets whose union contains U. That is, C must satisfy

$$U \subseteq \bigcup_{V \in C} V \qquad \text{and} \qquad \forall V \in C : V \in \mathcal{O}_{\mathcal{M}} \tag{2.15}$$

to constitute an open cover of U. A subcover $S \subseteq C$ is then a subset of C which still covers all of U.

The terminology of covers allows one to succinctly define compactness: A topological space $(\mathcal{M}, \mathcal{O})$ is compact if every open cover has a finite subcover, i.e. a subcover with a finite number of elements.

Given a metric space with the metric-induced topology (see section 2.2), it can be shown that this notion of compactness exactly coincides with the criterion given in the Heine-Borel theorem from

[6] Armed with this formal definition, it is not hard to recognise that the product topology $\mathcal{O}_{\mathcal{M} \times \mathcal{N}}$ coincides with the initial topology induced by the projection maps $\pi_{\mathcal{M}} : \mathcal{M} \times \mathcal{N} \longrightarrow \mathcal{M}$ and $\pi_{\mathcal{N}} : \mathcal{M} \times \mathcal{N} \longrightarrow \mathcal{N}$.

standard analysis which states that every closed and bounded set is compact. A slightly weaker property than compactness is given by paracompactness, which states that every open cover has a subcover which is locally finite, i.e. every point of the underlying space is only contained in finitely many elements of the subcover.

Many further properties of a space (which are not discussed in the scope of this thesis) are induced solely by the choice of a topology. Some of these topological properties are not only binary, such as connectedness, compactness, Hausdorff and so on but may instead even be group-valued, such as the fundamental group of a topological space.

However, no exhaustive list of such properties is known which could be used to classify the category of topological spaces. That is, while one can conclude that no homeomorphism can exist between two topological spaces which differ in any known topological property, the converse is not true: even if two topological spaces coincide with respect every known topological property, this is still not sufficient to conclude that they are homeomorphic.

2.2 Metric Spaces

A metric space (not to be confused with a metric manifold, see section 2.5) is a tuple $(\mathcal{M}, d)$ consisting of a set $\mathcal{M}$ and a function $d : \mathcal{M} \times \mathcal{M} \longrightarrow \mathbb{R}$ which satisfies for all points $x, y, z \in \mathcal{M}$:

(Positivity)	$0 \leq d(x, y)$	(2.16)
(Symmetry)	$d(x, y) = d(y, x)$	(2.17)
(Definiteness)	$d(x, y) = 0 \iff x = y$	(2.18)
(Triangle inequality)	$d(x, z) \leq d(x, y) + d(y, z).$	(2.19)

The function d is then said to be a metric on $\mathcal{M}$ or more informally a distance function. Note that the crucial requirement is the triangle inequality, since it captures the essence of the characteristic behaviour of a distance function. Namely, it encodes the fact that the direct path between two points should always be the shortest and that detours over any other intermediate point increase the distance, unless the intermediate point is already on the direct path.

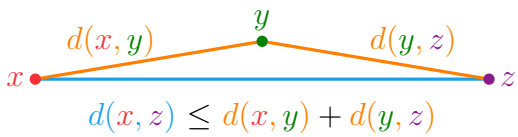


Figure 3: Visualisation of the triangle inequality (2.19).

A simple example of a metric space is given by the real line $\mathbb{R}$ equipped with the metric $d(x, y) = |x - y|$ which is constructed from the absolute value function.[7]

[7] Since $(\mathbb{R}, |\cdot|)$ constitutes a normed vector space, the function $d(x, y) = |x - y|$ is precisely the norm-induced metric.

Importantly, a metric d on a space $(\mathcal{M}, d)$ can be used to induce the so-called metric topology $\mathcal{O}_d$ on $\mathcal{M}$ via the open balls it generates

$$B_r(p) = \Big\{ q \in \mathcal{M} \;\Big|\; d(p,q) < r \Big\} \tag{2.20}$$

$$\mathcal{O}_d = \Big\{ U \subseteq \mathcal{M} \;\Big|\; \forall p \in U : \exists r > 0 : B_r(p) \subseteq U \Big\}. \tag{2.21}$$

Notably, metric topologies are always Hausdorff.

Given a function $h : X \longrightarrow Y$ that injectively maps from a space X into a metric space (Y, d_Y), one can induce a metric on X by

$$d_X(x,y) = d_Y\Big(h(x), h(y)\Big). \tag{2.22}$$

Clearly, d_X is positive and symmetric due to the positivity and symmetry of d_Y. The triangle inequality for d_X also follows from the triangle inequality for d_Y. It then remains to be shown that d_X is definite. Since h is injective by assumption, one has $h(x) = h(y) \iff x = y$. Thus for all $x, y \in X$ it is true that

$$0 = d_X(x,y) = d_Y\Big(h(x), h(y)\Big) \overset{d_Y \text{ definite}}{\Longleftrightarrow} h(x) = h(y) \overset{h \text{ injective}}{\Longleftrightarrow} x = y \tag{2.23}$$

which proves d_X is a metric. Thus, any space that can be injectively mapped into a metric space is itself metrisable.[8]

2.3 Vector Spaces

Before moving on to differentiable manifolds, which are supposed to be a generalisation of vector spaces, it is appropriate to first review vector spaces in their own right and recall all subtleties of the involved definitions.

For sake of increased conceptual clarity, different operations will be explicitly highlighted as such throughout this section, at the cost of slightly decreased readability. Also, elements of vector spaces are indicated by vector arrows here. However, considering that the truth is invariant under changes of notation, this over-explicit notation will soon be dropped in favour of the standard plain notation, where the exact meaning and properties of different operations and objects must be inferred from context.

An algebraic number field $(\mathbb{K}, +, \cdot)$ is a set $\mathbb{K}$ equipped with two maps

$$+ : \mathbb{K} \times \mathbb{K} \longrightarrow \mathbb{K}, \qquad \cdot : \mathbb{K} \times \mathbb{K} \longrightarrow \mathbb{K} \tag{2.24}$$

referred to as addition and multiplication that both separately satisfy commutativity, associativity, and that there exist neutral and inverse elements (**CANI**) for both operations respectively.[9]

[8] Since one essentially performs a pull-back of the metric on Y to X via the injective map h, one might also write this as $d_X = h^* d_Y$. See also section 2.8.

[9] In the case of multiplication, inverses must only exist for $\mathbb{K} \setminus \{0\}$ where 0 is the neutral element of addition.

Prominent examples of number fields are the real numbers $\mathbb{R}$ and the complex numbers $\mathbb{C}$.

A vector space V over a number field $\mathbb{K}$ is a triple $(V, \oplus, \odot)$ consisting of a set V and two operations $\oplus\colon V \times V \longrightarrow V$ and $\odot\colon \mathbb{K} \times V \longrightarrow V$ satisfying for all $\vec{u}, \vec{v}, \vec{w} \in V$ and $a, b \in \mathbb{K}$

(**C**ommutativity)		$\vec{u} \oplus \vec{v} = \vec{v} \oplus \vec{u}$	(2.25)
(**A**ssociativity)		$(\vec{u} \oplus \vec{v}) \oplus \vec{w} = \vec{u} \oplus (\vec{v} \oplus \vec{w})$	(2.26)
(**N**eutral element)	$\exists \vec{0} \in V\colon$	$\vec{v} \oplus \vec{0} = \vec{v}$	(2.27)
(**I**nverse element)	$\exists (-\vec{v}) \in V\colon$	$\vec{v} \oplus (-\vec{v}) = \vec{0}$	(2.28)
(**A**ssociativity)		$a \odot (b \odot \vec{v}) = (a \cdot b) \odot \vec{v}$	(2.29)
(**D**istributivity)		$a \odot (\vec{u} \oplus \vec{v}) = (a \odot \vec{u}) \oplus (a \odot \vec{v})$	(2.30)
(**D**istributivity)		$(a + b) \odot \vec{v} = (a \odot \vec{v}) \oplus (b \odot \vec{v})$	(2.31)
(**U**nit Element)	$\exists 1 \in \mathbb{K}\colon$	$1 \odot \vec{v} = \vec{v}.$	(2.32)

Crucially, vector spaces are not limited to collections of ordered tuples of real numbers which can be visualised by arrows. More generally, vector spaces can also be constituted by sets of functions, matrices and many other objects. Important examples of $\mathbb{R}$-vector spaces include for example the set of smooth real functions $C^\infty(\mathbb{R})$ and the smooth sections of the tangent or cotangent bundles $\Gamma(T\mathcal{M})$ and $\Gamma(T^*\mathcal{M})$ of a smooth manifold $\mathcal{M}$.

A finite subset $B = \{\vec{e}_1, ..., \vec{e}_d\} \subset V$ of a $\mathbb{K}$-vector space $(V, \oplus, \odot)$ is called a basis if

$$\forall \vec{v} \in V : \exists!\, v^1, ..., v^d \in \mathbb{K} : \qquad \vec{v} = \sum_{j=1}^{d} v^j \odot \vec{e}_j \tag{2.33}$$

that is, a unique assignment of components v^i with respect to the prospective basis B must exist for every vector. If such basis B exists, then the dimension of the vector space is defined to be $\dim V = |B|$.[10] Important instances of bases are orthogonal bases, where each pair of basis vectors is mutually orthogonal with respect to an inner product, and orthonormal bases, which are orthogonal bases where each basis vector has unit length with respect to the norm induced by the inner product.

Every vector space V also has a dual vector space denoted

$$V^* := \{L : V \longrightarrow \mathbb{K} \mid L \text{ linear}\} \tag{2.34}$$

which is the space of all linear maps from V into the underlying number field. Since linear maps can still be composed via addition and scalar multiplication, V^* is also a vector space. Informally, elements of a vector space V are referred to as "vectors" while elements of the dual vector space V^* are called "covectors". Importantly, for finite-dimensional vector spaces, one has $(V^*)^* \cong_{\text{vec.}} V$,

[10] Other notions of bases for vector spaces exist. While the basis presented here is a so-called Hamel basis for finite vector spaces, for infinite-dimensional vector spaces one uses the so called Schauder basis.

which means that a vector space and the dual of its dual space are isomorphic.

Having established such a basis $\{\vec{e}_1, ..., \vec{e}_d\}$ on V, one usually chooses the corresponding basis $\left\{\vec{\varepsilon}^1, ..., \vec{\varepsilon}^d\right\}$ on the dual space V^* such that they satisfy

$$\vec{\varepsilon}^a(\vec{e}_b) = \delta^a_b = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases} \tag{2.35}$$

where it should be stressed that no inner product or vector multiplication of any kind is involved. Instead, the linear maps $\vec{\varepsilon}^a \in V^*$ simply project the elements $\vec{e}_b \in V$ to the underlying number field $\mathbb{K}$ as per their definition.

2.3.1 Normed Vector Spaces

Norms impart a notion of length on the vector space. A normed vector space $(V, \|\cdot\|)$ is a vector space V equipped with a function $\|\cdot\| : V \longrightarrow \mathbb{K}$ satisfying for all $\vec{x}, \vec{y} \in V$ and $\lambda \in \mathbb{K}$

(Positivity)	$0 \leq \|\vec{x}\|$	(2.36)
(Definiteness)	$\|\vec{x}\| = 0 \iff \vec{x} = \vec{0}$	(2.37)
(Homogeneity)	$\|\lambda\vec{x}\| = \|\lambda\|\|\vec{x}\|$	(2.38)
(Triangle inequality)	$\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|.$	(2.39)

A norm on a vector space in turn naturally defines a metric on the space via $d(\vec{x}, \vec{y}) = \|\vec{x} - \vec{y}\|$. In cases where the inner product is induced by a positive definite matrix $B = A^\top A$, the metric given by

$$d^{(B)}(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^\top B(\vec{x} - \vec{y})} = \sqrt{(\vec{x} - \vec{y})^\top A^\top A\,(\vec{x} - \vec{y})} = d(A\vec{x}, A\vec{y}) \tag{2.40}$$

can be interpreted as the standard euclidean metric on a vector space transformed by A.

2.3.2 Inner Product Spaces

An inner product gives a natural way of defining the angle between two vectors through the ability of projecting vectors onto each other. In particular, it gives rise to the notion of orthogonality of vectors.

An inner product space is a vector space V over a number field $\mathbb{K}$ equipped with a function $\langle\cdot, \cdot\rangle : V \times V \longrightarrow \mathbb{K}$ which satisfies for all $\vec{x}, \vec{y}, \vec{z} \in V$ and $\lambda \in \mathbb{K}$

(Positivity)	$0 \leq \langle\vec{x}, \vec{x}\rangle$	(2.41)
(Symmetry)	$\langle\vec{x}, \vec{y}\rangle = \langle\vec{y}, \vec{x}\rangle$	(2.42)
(Definiteness)	$\langle\vec{x}, \vec{x}\rangle = 0 \iff \vec{x} = \vec{0}$	(2.43)
(Linearity in 1st slot)	$\langle\lambda\vec{x} + \vec{y}, \vec{z}\rangle = \lambda\langle\vec{x}, \vec{z}\rangle + \langle\vec{y}, \vec{z}\rangle.$	(2.44)

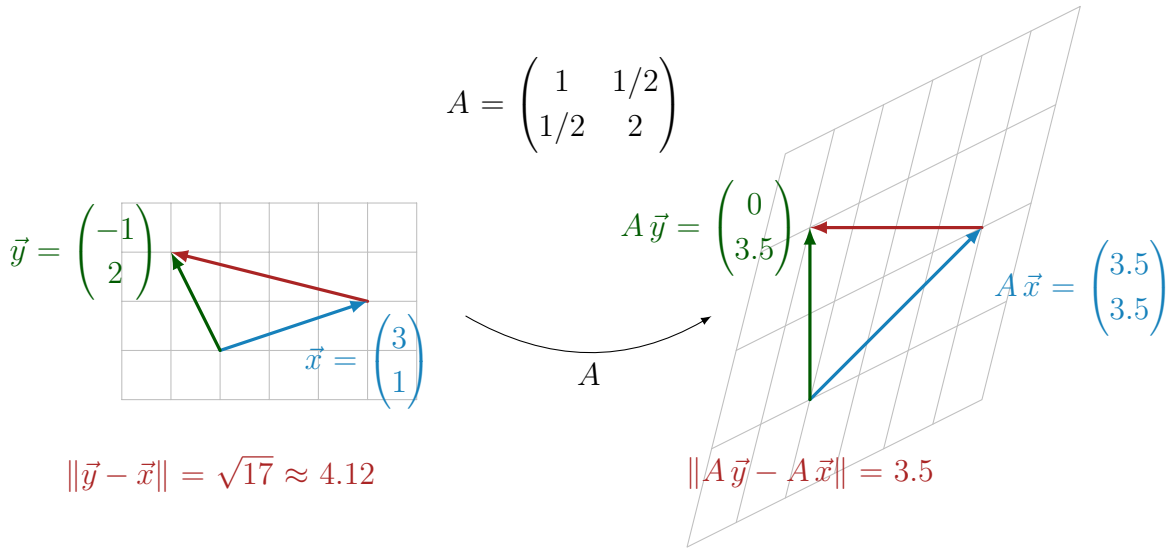


Figure 4: Visualisation of two-dimensional vectors under a linear transformation induced by a positive definite matrix A.

For vector spaces over number fields other than $\mathbb{R}$, the symmetry condition must often be extended. For $\mathbb{K} = \mathbb{C}$, the symmetry condition becomes the hermiticity condition: $\langle \vec{x}, \vec{y} \rangle = \overline{\langle \vec{y}, \vec{x} \rangle}$.[11] This hermiticity together with linearity in only one of the slots results in semilinearity for the other slot. That is, factors pulled out of the inner product from the second slot are complex conjugated. The resulting inner product is then said to be sesquilinear.

The standard inner product on the vector space $\mathbb{R}^n$ is given by

$$\langle \vec{x}, \vec{y} \rangle \equiv \vec{x} \cdot \vec{y} \coloneqq \sum_{j=1}^{n} x^j\, y^j = \vec{x}^{\top}\, \vec{y} \tag{2.45}$$

where $x^j \in \mathbb{R}$ and $y^j \in \mathbb{R}$ denote the components of the vectors $\vec{x}$ and $\vec{y}$ with respect to some basis. Two vectors $\vec{x}$ and $\vec{y}$ are then said to be orthogonal with respect to the inner product if they satisfy $\langle \vec{x}, \vec{y} \rangle = 0$.

However, many other choices of inner products are possible. For instance, consider the case that both entries of the scalar product are linearly transformed according to a positive definite matrix A before their inner product is taken, which can be expressed as

$$(A\vec{x}) \cdot (A\vec{y}) = (A\vec{x})^{\top}\, (A\vec{y}) = \vec{x}^{\top}\, \underbrace{A^{\top}\, A}_{=:B}\, \vec{y} \tag{2.46}$$

in some basis. One can check that a new inner product induced by a positive definite matrix B

$$\langle \vec{x}, \vec{y} \rangle_B \coloneqq \vec{x}^{\top}\, B\, \vec{y} \tag{2.47}$$

fulfils the necessary requirements for a vector space $\mathbb{R}^n$. In particular, for the positive definite matrix A from the previous example, the scalar product induced by $B = A^{\top}\, A$ can now be

[11] Here, $\overline{z}$ denotes the complex conjugate of z.

given the interpretation of constituting the standard inner product on a space in which has been transformed according to A. Conversely, it is also true that any positive definite matrix B has a unique Cholesky decomposition, which means that it can be uniquely factorised into a product $A^\top A$ where A is an upper triangular matrix.

From this, one can conclude that any choice of inner product on a vector space V which can be represented by a positive definite matrix $B = A^\top A$ is equivalent to the choosing the standard inner product. That is, there exists an isomorphism between the two inner product spaces which is precisely given by A^{-1}. This isomorphism not only preserves the vector space structure but also preserves the inner product by mapping orthonormal bases from one space to the other. This result is closely related to the well-known Sylvester's theorem.

An example of an (infinite-dimensional) inner product space is given by the space of real continuous functions over an interval $C^0([a,b])$, equipped with the integral

$$\langle f, g\rangle \coloneqq \int_a^b \mathrm{d}x\, f(x)\, g(x) \tag{2.48}$$

as its inner product. Inner products induce a natural norm via the definition

$$\|\vec{x}\| \coloneqq \sqrt{\langle \vec{x}, \vec{x}\rangle}. \tag{2.49}$$

An important inequality for both inner products and their induced norms is the Cauchy–Schwarz inequality which states for all $\vec{x}, \vec{y} \in V$

$$\left|\langle \vec{x}, \vec{y}\rangle\right|^2 \leq \langle \vec{x}, \vec{x}\rangle \langle \vec{y}, \vec{y}\rangle \tag{2.50}$$

from which it immediately follows that

$$\left|\langle \vec{x}, \vec{y}\rangle\right| \leq \|\vec{x}\| \|\vec{y}\| \tag{2.51}$$

by application of the square root on both sides. In particular, the Cauchy–Schwarz inequality becomes an equality if the arguments $\vec{x}$ and $\vec{y}$ are linearly dependent. A detailed proof and discussion of the Cauchy–Schwarz inequality can be found in [32].

2.3.3 Tensor Spaces

It is hard to overstate the importance of tensors in modern fundamental physics. Since they allow for a chart-independent description, they lend themselves particularly well for modelling physical quantities such as the electromagnetic field strength, the curvature of spacetime and so on.

Given a vector space V over some number field, say, the real numbers $\mathbb{R}$, then the space of multi-linear maps defined by

$$T^{(p,q)}V \coloneqq \left\{ S : \underbrace{V^* \times \ldots \times V^*}_{p} \times \underbrace{V \times \ldots \times V}_{q} \longrightarrow \mathbb{R} \;\middle|\; S \text{ multi-linear} \right\} \tag{2.52}$$

is called the tensor space $T^{(p,q)}V$, where multi-linearity refers to the fact that the maps ought to be linear in each of their slots. Moreover, this tensor space is equipped with pointwise addition $\boxplus : T^{(p,q)}V \times T^{(p,q)}V \longrightarrow T^{(p,q)}V$ and multiplication $\boxdot : \mathbb{R} \times T^{(p,q)}V \longrightarrow T^{(p,q)}V$ such that as a whole, $(T^{(p,q)}V, \boxplus, \boxdot)$ is a vector space in its own right. Furthermore, there is a tensor product which allows for combination of tensors of different valence[12] given by

$$\otimes : T^{(p,q)}V \times T^{(r,s)}V \longrightarrow T^{(p+r,q+s)}V. \tag{2.53}$$

For example, the tensor product of two tensors $S, Q \in T^{(1,1)}V$ is constructed by

$$(S \otimes Q)(u, v, \omega, \sigma) \coloneqq S(u, \omega) \cdot Q(v, \sigma). \tag{2.54}$$

for all $u, v \in V$ and $\omega, \sigma \in V^*$. As for any vector space, one may of course choose a basis with respect to which tensors can be described in components. Specifically, the components of a tensor $S \in T^{(p,q)}V$ are found by evaluating tensors on the basis vectors of the underlying vector spaces V and V^*

$$S^{i_1 \dots i_p}{}_{j_1 \dots j_q} \coloneqq S\Big(\varepsilon^{i_1}, \dots, \varepsilon^{i_p}, e_{j_1}, \dots, e_{j_q}\Big). \tag{2.55}$$

Due to the multi-linearity of tensors, the components can be used to recover the original tensor via

$$S = \sum_{i_1=1}^{\dim V} \dots \sum_{i_p=1}^{\dim V} \sum_{j_1=1}^{\dim V} \dots \sum_{j_q=1}^{\dim V} S^{i_1 \dots i_p}{}_{j_1 \dots j_q}\, e_{i_1} \otimes \dots \otimes e_{i_p} \otimes \varepsilon^{j_1} \otimes \dots \otimes \varepsilon^{j_q}. \tag{2.56}$$

Especially for tensors of high rank, writing out all of the above sums over indices explicitly tends to clutter up the notation. Therefore, most authors adopt the Einstein summation convention, which consists of the tacit agreement that whenever the same index variable appears both in the contravariant position (i.e. as a superscript) and in the covariant position (i.e. as a subscript) within the same summand, it is implied that this term should be summed over all values of the index. Likewise, the Einstein summation convention is employed from this point onwards. As indicated by equations (2.55) and (2.56), not only the order of indices but also whether they are in the contravariant or covariant position plays a crucial role when using the outlined index notation.

Because the basis of a tensor space $T^{(p,q)}V$ is usually constructed from tensor products of basis elements from the underlying vector space V, a change of basis on V will also lead to a transformation of tensor components. In short, a linear transformation $L : V \longrightarrow V$ defined by $L(v)^i \coloneqq L^i{}_j\, v^j = \tilde{v}^i$ leads to a transformation behaviour for the components of a tensor $S \in T^{(1,3)}V$ according to

$$\tilde{S}^a{}_{bc} = L^a{}_i\, (L^{-1})^j{}_b\, (L^{-1})^k{}_c\, S^i{}_{jk} \tag{2.57}$$

and analogously for tensors of other valence. Since transformations of tensors only result in multiplicative factors, this implies that if a tensor vanishes in one coordinate system, it must vanish in all coordinate system.

[12] The tuple (p, q) of a tensor space $T^{(p,q)}V$ is referred to as the valence of a tensor while the sum $p + q$ is called the rank. However, it should be noted that these terms are not used consistently by all authors throughout the literature.

Important subclasses of tensors are symmetric and antisymmetric tensors.[13] As the name suggests, they are respectively characterised by the fact that any odd permutation in their arguments either leaves the value of the tensor invariant or results in a minus sign. Specifically, for two tensors $S, A \in T^{(0,2)}V$ where S is symmetric and A is antisymmetric, one has for any two vectors $u, v \in V$

$$S(u,v) = S(v,u) \qquad \Longleftrightarrow \qquad S_{ab} = S_{ba} \tag{2.58}$$
$$A(u,v) = -A(v,u) \qquad \Longleftrightarrow \qquad A_{ab} = -A_{ba} \tag{2.59}$$

and similarly for tensors of higher rank. A convenient and widely-used notation to denote the symmetrised and antisymmetrised parts of a tensor $Q \in T^{(0,2)}V$ in components is by enveloping the symmetrised and antisymmetrised indices using round and square brackets respectively, i.e.

$$Q_{(ab)} \coloneqq \frac{1}{n!}(Q_{ab} + Q_{ba}) \qquad \text{and} \qquad Q_{[ab]} \coloneqq \frac{1}{n!}(Q_{ab} - Q_{ba}) \tag{2.60}$$

where n is the number of indices which are symmetrised or antisymmetrised, i.e. $n = 2$ here. Moreover, any tensor $Q \in T^{(0,2)}V$ can be uniquely decomposed into its symmetric and antisymmetric parts as

$$Q_{ab} = Q_{(ab)} + Q_{[ab]}. \tag{2.61}$$

For the notation to remain consistent, one should denote the $\binom{p}{q}$-tensor space over the tangent space $T_x\mathcal{M}$ by $T^{(p,q)}(T_x\mathcal{M})$, however, the notation is usually shortened to $T_x^{(p,q)}\mathcal{M}$ since there is no ambiguity about the underlying vector space. Similarly, the set of smooth sections of a tangent bundle (see section 2.7) is generally abbreviated as $\Gamma(T\mathcal{M})$ rather than the technically more accurate $\Gamma\big(T^{(1,0)}(T\mathcal{M})\big)$. For better or worse, this common abbreviation is also adopted here.

2.4 Manifolds

In essence, differential geometry is the study of coordinate transformations and their effects on various mathematical objects. Specifically, it aims to provide a language which expressly focuses on properties which remain invariant under smooth changes of coordinates. Similar to how multi-variable calculus is an extension of the one-dimensional calculus one learns in high school, differential geometry aims to again lift the same concepts from vector spaces to a more general class of spaces that is only locally isomorphic to a vector space.

An example of this is given by the torus in figure 5. Roughly speaking, the torus does not constitute a vector space because there is no consistent way of adding and scaling points on its surface such that one always obtains another point on the torus. On the other hand, if one sufficiently zooms into the surface of a torus, it approximately looks like a flat two-dimensional plane. The same is

[13] In fact, totally antisymmetric covariant tensor fields over a smooth manifold give rise to the so-called exterior algebra, which naturally features a rich set of operations such as the gradient, wedge product and the Hodge dual without the need of introducing any further structure on the manifold.

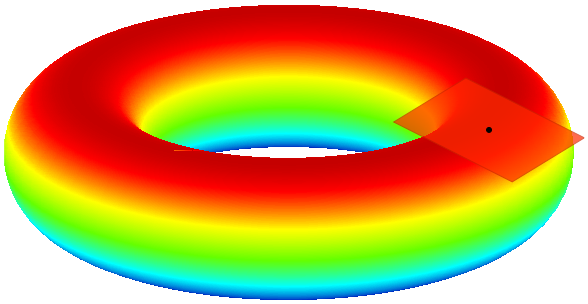


Figure 5: Illustration of a torus which is locally homeomorphic to $\mathbb{R}^2$ and can therefore be made into a smooth manifold.

also true for the surface of a sphere.

2.4.1 Charts, Atlases and Differentiable Structures

A paracompact Hausdorff topological space $(\mathcal{M}, \mathcal{O})$ is said to form an n-dimensional topological manifold if it is locally homeomorphic to $(\mathbb{R}^n, \mathcal{O}_{\text{std}})$ for some $n \in \mathbb{N} \setminus \{0\}$. That is, the topological space $(\mathcal{M}, \mathcal{O})$ must formally satisfy

$$\forall p \in \mathcal{M} \colon \exists U \in \mathcal{O} \colon p \in U \colon \qquad \exists x : U \longrightarrow x(U) \subseteq \mathbb{R}^n \tag{2.62}$$

where the map x is invertible and continuous both ways.

A tuple (U, x) where $U \in \mathcal{O}$ is an open set in $\mathcal{M}$ and $x : U \longrightarrow x(U) \subseteq \mathbb{R}^n$ is a homeomorphic map into $\mathbb{R}^n$ is referred to as a chart of the topological space $(\mathcal{M}, \mathcal{O})$. A collection of charts is then called an atlas. In particular, the set of all possible charts on a topological space $(\mathcal{M}, \mathcal{O})$ is said to form the maximal atlas $\mathcal{A}_{\text{max}}$ of $(\mathcal{M}, \mathcal{O})$ defined by

$$\mathcal{A}_{\text{max}} := \Big\{ (U, x) \;\Big|\; U \in \mathcal{O}, x : U \longrightarrow \mathbb{R}^n \text{ is a homeomorphism} \Big\} . \tag{2.63}$$

Since the maximal atlas does not involve any choice or condition beyond the specification of a topological space $(\mathcal{M}, \mathcal{O})$, a topological manifold can be seen as already implicitly equipped with its maximal atlas, i.e. it forms a triple $(\mathcal{M}, \mathcal{O}, \mathcal{A}_{\text{max}})$.

Given a topological manifold $(\mathcal{M}, \mathcal{O}, \mathcal{A}_{\text{max}})$, a so-called differentiable structure can be established on it by systematically removing charts from the maximal atlas in such a way that all the remaining chart transition maps satisfy some differentiability condition in addition to being homeomorphisms. Since the desired differentiability condition may differ depending on the application one has in mind, the symbol ☕ is used as a placeholder for this condition in the following.

Two charts $(U, x), (V, y) \in \mathcal{A}$ are said to be ☕-compatible if

$$U \cap V \neq \varnothing \qquad \Longrightarrow \qquad \begin{array}{ll} (y \circ x^{-1}) : x(U \cap V) \longrightarrow y(U \cap V) & \text{and its inverse} \\ (x \circ y^{-1}) : y(U \cap V) \longrightarrow x(U \cap V) & \text{both satisfy ☕.} \end{array}$$

where $x(U \cap V), y(U \cap V) \subseteq \mathbb{R}^{\dim \mathcal{M}}$. If $U \cap V = \varnothing$, the condition ☕ is trivially fulfilled. An illustration of chart transition maps of a topological manifold is given in figure 6.

For example, ☕ can denote C^k or C^∞, i.e. k-fold continuous differentiability or smoothness. On the other hand, ☕ can also denote real analytic or even complex differentiable and so on. In every case, it is understood that the condition must apply to all possible chart transition maps within the same atlas. A topological space equipped with a ☕-atlas is then called a ☕ manifold.

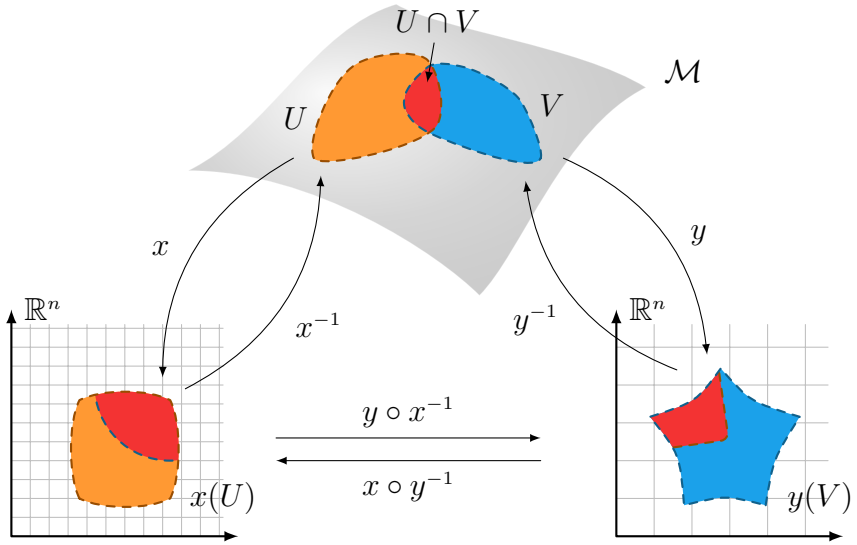


Figure 6: Visual representation of overlapping charts and their transition maps on a topological manifold $(\mathcal{M}, \mathcal{O}, \mathcal{A}_{\text{max}})$ that is locally homeomorphic to $(\mathbb{R}^n, \mathcal{O}_{\text{std}})$. This illustrates that effects observed in a chart such as distortions do not necessarily correspond to features on the underlying manifold but may in fact be artefacts of the chosen chart representation.

In differential geometry, one usually considers smooth manifolds, since this is more convenient and only marginally less general than considering k-times differentiable manifolds. This is due a theorem by Whitney which states that any maximal C^k-atlas for $k \geq 1$ contains a C^∞-atlas. Moreover, the theorem affirms that two maximal C^k-atlases which contain the same C^∞-atlas are already identical. This shows that there is no real need to distinguish between a C^k-manifolds with $k \geq 1$ and smooth manifolds as it is always possible to further restrict a C^k-atlas until it is smooth.

To illustrate how two differentiable structures on a manifold can be incompatible, consider the real line equipped with the standard topology $(\mathbb{R}, \mathcal{O}_{\text{std}})$ and two possible differential structures constituted by the atlases $\mathcal{A}_1 = \left\{(\mathbb{R}, x(\alpha) = \alpha^3)\right\}$ and $\mathcal{A}_2 = \left\{(\mathbb{R}, y(\alpha) = \alpha)\right\}$. Since both atlases only contain a single chart each, differentiability conditions on the chart transition maps (which in both cases is the identity map on $\mathbb{R}$) are trivially satisfied.

However, if one considers the union $\mathcal{A}_1 \cup \mathcal{A}_2 = \left\{(\mathbb{R}, x), (\mathbb{R}, y)\right\}$ one finds that the chart transition map

$$\left(y \circ x^{-1}\right)(\alpha) = \alpha^{\frac{1}{3}} \tag{2.64}$$

is not continuously differentiable at $\alpha = 0$ and thus the two atlases are incompatible.

Once suitable differentiable structures are in place, one can lift the notion of a smooth map from the chart level to the manifold level: a map $h : \mathcal{M} \longrightarrow \mathcal{N}$ between two smooth manifolds $\mathcal{M}$ and $\mathcal{N}$ is said to be smooth if the local representations $y \circ h \circ x^{-1}$ of h in all possible pairs of charts $(U, x) \in \mathcal{A}_\mathcal{M}$, $(V, y) \in \mathcal{A}_\mathcal{N}$ are smooth as functions from $\mathbb{R}^{\dim \mathcal{M}}$ to $\mathbb{R}^{\dim \mathcal{N}}$ (in the sense of standard real analysis). Figure 7 demonstrates that it suffices to show that there exists at least one set of smooth local chart representations of h which covers all of $\mathcal{M}$ and $\mathcal{N}$ to conclude that h is smooth as a map between manifolds.[14]

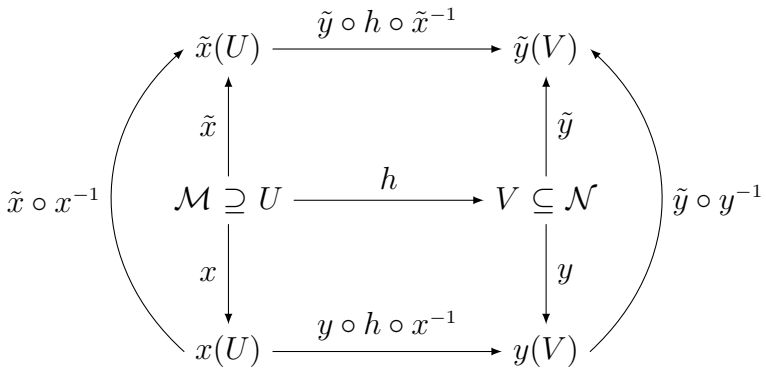


Figure 7: This graph depicts a map $h : \mathcal{M} \longrightarrow \mathcal{N}$ which relates two subdomains $U \subseteq \mathcal{M}$ and $V \subseteq \mathcal{N}$ with $\mathcal{M}$ and $\mathcal{N}$ smooth manifolds. To check whether h is smooth as a map between manifolds, one can see from the graph that it suffices to check whether it is smooth in the sense of standard analysis (i.e. as a map from $\mathbb{R}^{\dim \mathcal{M}}$ to $\mathbb{R}^{\dim \mathcal{N}}$) with respect to any single pair of charts $(U, x) \in \mathcal{A}_\mathcal{M}$ and $(V, y) \in \mathcal{A}_\mathcal{N}$. Since compositions of smooth maps are again smooth, any change of chart on the domain or target preserves the smoothness (or non-smoothness) of the chart representation of $y \circ h \circ x^{-1}$ that was checked.

A map $h : \mathcal{M} \longrightarrow \mathcal{N}$ which is invertible and smooth both ways is referred to as a diffeomorphism. Similar to homeomorphisms, which constitute isomorphisms between topological spaces, diffeomorphisms are isomorphisms (i.e. structure-preserving maps) between smooth manifolds.

A question that naturally arises is how many inequivalent differentiable structures can be established on a given topological space. The Radon–Moise theoreoms state that for $1 \leq \dim \mathcal{M} \leq 3$, there is a unique smooth manifold (up to diffeomorphisms) that can be build from any topological manifold $(\mathcal{M}, \mathcal{O}, \mathcal{A}_{\text{max}})$. For compact topological manifolds $(\mathcal{M}, \mathcal{O}, \mathcal{A}_{\text{max}})$ with $\dim \mathcal{M} > 4$, it was shown that the number of inequivalent smooth structures (up to diffeomorphisms) is finite. However, no such bound is known for topological manifolds with $\dim \mathcal{M} = 4$. Moreover, there are counterexamples of topological manifolds on which uncountably many inequivalent smooth structures can be established.

Unless specified otherwise, it is understood that $\mathbb{R}$ is equipped with the standard topology $\mathcal{O}_{\text{std}}$ and the atlas which only contains the identity chart $\mathcal{A}_{\text{std}} = \big\{(\mathbb{R}, \text{id})\big\}$, whenever such structures

[14] Conversely, the existence of any local chart representation of h which is not smooth implies that h is decidedly not smooth as a map between manifolds.

are required.[15] To lighten the notation, the precise structures on $\mathbb{R}$ will often be suppressed with the tacit agreement that it is always equipped with the least amount of structure necessary.

2.4.2 Tangent and Cotangent Spaces

Tangent and cotangent spaces play a pivotal role differential geometry. Since they locally approximate the manifold as a real vector space, they constitute an essential tool in generalising concepts such as instantaneous velocity, length and angles from linear algebra to smooth manifolds.

As such, there exist at least 3 equivalent ways to define the notion of a tangent space:

1. through the desired transformation behaviour of its elements under changes of coordinates,
2. via the action of derivations on smooth functions, i.e. algebraically,
3. by defining vectors as instantaneous velocities of curves, i.e. geometrically.

Since it is arguably the most intuitive approach of the three, the geometric definition of tangent spaces is outlined in the following.

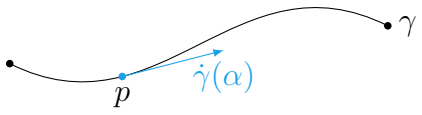


Figure 8: Visualisation of a tangent vector to a smooth curve γ in $\mathbb{R}^2$ with $\gamma(\alpha) = p$. Although the tangent vector $\dot{\gamma}(\alpha)$ is technically an element of $T_p\mathbb{R}^2$ and therefore not a part of the space $\mathbb{R}^2$ in which the curve lies, the two spaces are isomorphic ($T_p\mathbb{R}^2 \cong \mathbb{R}^2$) wherefore this picture nonetheless gives the right idea.

Given a smooth curve[16] $\gamma : \mathbb{R} \longrightarrow \mathcal{M}$ one can find its tangent vector at a particular parameter value, also referred to as the velocity of the curve, by differentiating with respect to the parameter. The dot and prime notations $\dot{\gamma} \equiv \gamma'$ are used interchangeably to indicate a derivative with respect to the parameter slot, whenever it is unambiguous which slot is meant, i.e. whenever a map only has a single slot. To perform the derivative explicitly in components, a parametrisation of the curve must be specified as $x \circ \gamma$ in some chart (U, x).

$$\frac{\mathrm{d}\gamma^a}{\mathrm{d}t} \equiv \dot{\gamma}^a \coloneqq \frac{\mathrm{d}}{\mathrm{d}t}(x^a \circ \gamma) \tag{2.65}$$

Next, consider the space of all smooth curves which pass through a point $p \in \mathcal{M}$

$$\kappa_p \coloneqq \left\{ \gamma : (-\epsilon, \epsilon) \longrightarrow \mathcal{M} \;\middle|\; \gamma(0) = p, \gamma \text{ smooth} \right\}. \tag{2.66}$$

[15] Furthermore, $\mathbb{R}$ is understood to carry the usual vector space structure $(\mathbb{R}, +, \cdot)$, the euclidean norm and its induced distance function (or even the Borel σ-algebra and the Lebesgue measure and so on) whenever necessary.

[16] As outlined in section 2.4.1, the smoothness of a map can only be determined between smooth manifolds. Accordingly, it is understood that $\mathbb{R} \mathrel{\widehat{=}} (\mathbb{R}, \mathcal{O}_{\text{std}}, \mathcal{A}_{\text{std}})$ in this case.

The tangent space $T_p\mathcal{M}$ is then given by the set of all tangent vectors of smooth curves through p, evaluated at p. That is,

$$T_p\mathcal{M} \coloneqq \left\{\dot{\gamma}(0) \;\middle|\; \gamma \in \kappa_p\right\} = \left\{X_{\gamma,p} \;\middle|\; \gamma \in \kappa_p\right\} \tag{2.67}$$

where elements $X \in T_p\mathcal{M}$ can be considered as linear maps $X : C^\infty(\mathcal{M}) \xrightarrow{\sim} \mathbb{R}$.[17] A common notation for the tangent vector to a curve γ at a point p is $X_{\gamma,p}$. Said tangent vector then acts on any $f \in C^\infty(\mathcal{M})$ by

$$X_{\gamma,p}\, f \coloneqq (f \circ \gamma)'(0) \tag{2.68}$$

which highlights that it can also be thought of as a directional derivative of f along γ at $p \in \mathcal{M}$.

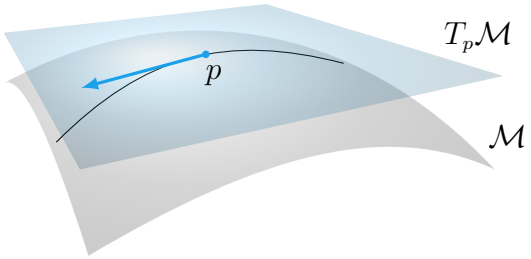


Figure 9: Illustration of a tangent space to a point in a smooth manifold which arises as the collection of all tangent vectors of all possible curves through the point.

It is not hard to see that elements of $T_p\mathcal{M}$ can be scaled arbitrarily by a factor $\alpha \in \mathbb{R}$ and again yield an element of $T_p\mathcal{M}$ by reparametrising the curve γ which generates them, i.e.

$$\eta(\lambda) = \gamma(\alpha \cdot \lambda) \qquad \Longrightarrow \qquad \dot{\eta}(0) = \alpha \odot \dot{\gamma}(0). \tag{2.69}$$

On the other hand, to show that $(T_p\mathcal{M}, \oplus, \odot)$ constitutes a vector space, it also has to be shown that it is closed under an addition $\oplus$ and that the "CANI ADDU" vector space axioms are satisfied (see section 2.3).

One can check that the pointwise definition of the operation $\oplus : T_p\mathcal{M} \times T_p\mathcal{M} \longrightarrow T_p\mathcal{M}$ via

$$\left(X_{\gamma,p} \oplus X_{\delta,p}\right)(f) \coloneqq X_{\gamma,p}(f) + X_{\delta,p}(f) \qquad \forall f \in C^\infty(\mathcal{M}) \tag{2.70}$$

is well-defined for any curves $\gamma, \delta \in \kappa_p$ and fulfils the necessary vector space axioms. Explicitly, one can construct the smooth curve $\sigma \in \kappa_p$ whose tangent vector at the point p is the sum of the tangent vectors $X_{\gamma,p}$ and $X_{\delta,p}$ by employing a chart (U, x) with $p \in U \subseteq \mathcal{M}$ as

$$\sigma(\lambda) = x^{-1} \circ \Big((x \circ \gamma)(\lambda) + (x \circ \delta)(\lambda) - x(p)\Big). \tag{2.71}$$

One can verify that this curve indeed generates the desired unique element of $T_p\mathcal{M}$ by letting it

[17] The linearity of a map can only be determined between vector spaces. Indeed, $\left(C^\infty(\mathcal{M}), +, \cdot\right)$ with its addition and scalar multiplication defined pointwise constitutes an infinite-dimensional vector space.

act on an arbitrary smooth function $f \in C^\infty(\mathcal{M})$

$$X_{\sigma,p}\, f = (f \circ \sigma)'(0) = \Big[\Big(f \circ x^{-1}\Big) \circ \Big(x \circ \gamma + x \circ \delta - x(p)\Big)\Big]'(0) \tag{2.72}$$

$$= \Big(f \circ \gamma + f \circ \delta - \underbrace{f(p)}_{\text{const.}}\Big)'(0) = (f \circ \gamma)'(0) + (f \circ \delta)'(0) = \Big(X_{\gamma,p} \oplus X_{\delta,p}\Big)(f). \tag{2.73}$$

Thus, the addition as well as the s-multiplication on $T_p\mathcal{M}$ are reduced to the addition and multiplication on $\mathbb{R}$ and consequently it is straightforward to see that they inherit the vector space structure from $\mathbb{R}$. At this point, it is important to emphasise that although one can scale tangent vectors arbitrarily, there is not yet any notion of length in place. Instead, it is only after the introduction of a metric tensor field on the manifold, that a notion of length can be defined independent of any choice of chart.

Usually, one chooses to induce the basis on $T_p\mathcal{M}$ from the choice of chart (U, x) on the underlying manifold by via $\left\{\frac{\partial}{\partial x^1}, ..., \frac{\partial}{\partial x^n}\right\}$ with $n = \dim \mathcal{M}$. The action of these basis vectors of $T_p\mathcal{M}$ on smooth functions $f \in C^\infty(\mathcal{M})$ is then given by

$$\frac{\partial}{\partial x^a}\bigg|_p f \coloneqq \Big(\partial_a (f \circ x^{-1})\Big)\Big(x(p)\Big) \tag{2.74}$$

where ∂_a denotes the partial derivative with respect to the a-th slot of $f \circ x^{-1}$. As discussed in section 2.3 this in turn induces a basis on the dual space $T_p^*\mathcal{M}$ which is given by $\left\{\mathrm{d}x^1, ..., \mathrm{d}x^n\right\}$.[18] Indeed, it is not hard to see that $\mathrm{d}x^a\Big(\frac{\partial}{\partial x^b}\Big) = \frac{\partial x^a}{\partial x^b} = \delta^a_b$.

As a result of the way manifolds were defined, the tangent spaces at every pair of points $p, q \in \mathcal{M}$ are isomorphic,[19] that is, $T_p\mathcal{M} \cong_{\text{vec.}} T_q\mathcal{M}$. However, despite this isomorphism, it is not possible to combine or compare vectors from different tangent spaces $p \neq q$ in a meaningful (chart-independent) way without establishing further structure on the manifold. Namely, this requires a so-called connection (see section 2.6.1) which induces a notion of parallel transport.

2.5 Metric Tensor Fields

The specification of a metric tensor field (often simply referred to as a metric but not to be confused with metric functions discussed in section 2.2) endows a manifold with a "rigid shape" through the notion of length. Moreover, it also gives rise to the notion of angles between tangent vectors by defining an inner product on the tangent spaces.

Concretely, a Riemannian metric g is a smooth $\binom{0}{2}$-tensor field, which satisfies the requirements of an inner product stated in equations (2.41) to (2.44) separately in every tangent space, i.e. as a map $g_p \colon T_p\mathcal{M} \times T_p\mathcal{M} \longrightarrow \mathbb{R}$ at every point $p \in \mathcal{M}$. The fact that this inner product changes smoothly across the manifold is encoded by the requirement that it be a smooth section of the $\binom{0}{2}$-tensor bundle over $\mathcal{M}$, that is, $g \in \Gamma(T^{(0,2)}\mathcal{M})$.

[18] The dual space of the tangent space at a point $p \in \mathcal{M}$ is generally denoted as $T_p^*\mathcal{M} \coloneqq (T_p\mathcal{M})^*$.
[19] In particular, this implies that dimension of the tangent spaces must be the same across the whole manifold.

For a $\binom{1}{1}$-tensor over some real finite-dimensional vector space V, it is always possible to find a basis in which its component matrix becomes diagonal and where said diagonal consists of a unique set of eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. However, for $\binom{2}{0}$-tensors or $\binom{0}{2}$-tensors, such a set of unique eigenvalues does not exist due to their differing transformation behaviour. Instead, a theorem due to Silvester guarantees that for such tensors there always exists a chart in which its component matrix is diagonal, but the diagonal is filled with values that are either $+1$, -1 or 0.

The triple (n_+, n_-, n_0) where n_+, n_- and n_0 respectively denote the total number of ones, minus ones and zeros on the diagonal is then called the signature. Importantly, this signature is unique for a given tensor.[20] If $n_0 = 0$, that is, if the diagonalised component matrix of a $\binom{0}{2}$-tensor does not contain any zeros, the signature is referred to as non-degenerate or pseudo-Riemannian. This effects that the tensor is invertible, which is a desirable property in almost all practical applications. A metric g is said to be Riemannian if it has the signature $(\dim \mathcal{M}, 0, 0)$.

Apart from Riemannian metrics, another important subclass of pseudo-Riemmanian metrics are Lorentzian metrics, which have the signature $(1, \dim \mathcal{M} - 1, 0)$. Specifically, space-time manifolds which are used to describe relativistic physics carry a Lorentzian signature of $(1, 3)$ which is necessary for compatibility with the well-tested Maxwell equations of electrodynamics. Notably, the requirement of positivity (2.41) is dropped and the definiteness property (2.43) is weakened to a non-degeneracy criterion for non-Riemannian metrics. As a result of this non-definiteness of Lorentzian metrics, a significant number of central theorems for Riemannian manifolds are no longer valid. An in-depth treatment of the subtle differences between Riemannian and Lorentzian manifolds can be found in [12].

The inverse metric $g_p^{-1} : T_p^*\mathcal{M} \times T_p^*\mathcal{M} \longrightarrow \mathbb{R}$ can be defined pointwise in components as

$$g_{ab}\,(g^{-1})^{bc} \stackrel{!}{=} \delta_a^c \tag{2.75}$$

which shows that its component matrix is the inverse of the component matrix of $g_p \colon T_p\mathcal{M} \times T_p\mathcal{M} \longrightarrow \mathbb{R}$ in the sense of linear algebra. The smoothness of the inverse metric tensor field g^{-1} is a direct consequence of the smoothness of g.

Due to the positivity condition (2.41), Riemannian metrics induce a norm given by

$$\|X\| \coloneqq \sqrt{g(X, X)} \tag{2.76}$$

exists for all vector fields X. By extension, the length of a curve $\gamma \colon (a, b) \longrightarrow \mathcal{M}$ is then given by

$$L[\gamma] \coloneqq \int_a^b \mathrm{d}t \left\|\dot{\gamma}(t)\right\| = \int_a^b \mathrm{d}t \sqrt{g_{\gamma(t)}\big(\dot{\gamma}(t), \dot{\gamma}(t)\big)} \tag{2.77}$$

where $\dot{\gamma}$ denotes the tangent vector to the curve γ. In effect, the length $L[\gamma]$ is measured by integrating up the lengths of all tangent vectors along the curve γ. Using this notion of length, one

[20] Clearly, this classification of $\binom{0}{2}$-tensors by their signature contains much less information than knowledge of the eigenvalues of a $\binom{1}{1}$-tensor due to the fact that whereas there are only a finite number of possible signatures, the number of realisable eigenvalue combinations is infinite.

can define the distance between two points $p, q \in \mathcal{M}$ via

$$d(p,q) \coloneqq \inf_{\{\gamma \,|\, \gamma(a)=p, \gamma(b)=q\}} L[\gamma] \tag{2.78}$$

which can be shown to fulfil the necessary requirements for a metric.

2.5.1 The Musical Isomorphism

A non-degenerate metric tensor field induces the so-called musical isomorphism, which identifies elements of the tangent and cotangent spaces at a point with one another. Colloquially, this is also referred to as raising and lowering indices of tensor components. The name and the notation used for this isomorphism are borrowed from music theory in that one denotes the process of associating a tangent space element to a given cotangent space element (i.e. raising the index) by $\sharp$ while the inverse of this bijective association is denoted using $\flat \coloneqq \sharp^{-1}$.

In a nutshell, the musical isomorphism is given by

$$\sharp : T_p^*\mathcal{M} \longrightarrow T_p\mathcal{M} \qquad \omega \longmapsto \sharp(\omega) \coloneqq g^{-1}(\omega, \cdot\,), \tag{2.79}$$

$$\flat : T_p\mathcal{M} \longrightarrow T_p^*\mathcal{M} \qquad Y \longmapsto \flat(Y) \coloneqq g(Y, \cdot\,). \tag{2.80}$$

Importantly, the multi-linearity of both the metric tensor and inverse metric tensor imply that the musical isomorphism facilitates $T_p\mathcal{M} \cong_{\text{vec.}} T_p^*\mathcal{M}$ for all $p \in \mathcal{M}$. Furthermore, the fact that the metric and inverse metric are both smooth tensor fields implies that both $\sharp$ and $\flat$ constitute diffeomorphisms, which means they change smoothly across the manifold. Thus, their existence also shows that the entire tangent and cotangent bundle are isomorphic as smooth manifolds which can be written as $T\mathcal{M} \cong_{\text{smooth}} T^*\mathcal{M}$.

Although it may seem like an abuse of this formalism, one can consider the natural contraction between tensor components as an inner product by first raising or lowering the index of one of the two objects and subsequently applying the metric-induced inner product on the tangent or cotangent space. For example, for $X \in \Gamma(T\mathcal{M})$ and $\omega \in \Gamma(T^*\mathcal{M})$, one may write

$$\omega(Y) = \omega_a\, Y^a = g\Big(\sharp(\omega), Y\Big) = g^{-1}\Big(\omega, \flat(Y)\Big). \tag{2.81}$$

In this way, one can exploit results like the Cauchy-Schwarz inequality for inner product spaces in the context of tensor contractions.

2.5.2 Geodesics on Manifolds

Given a smooth manifold $(\mathcal{M}, \mathcal{O}, \mathcal{A})$ which is equipped with a Riemannian metric g, the geodesics on $\mathcal{M}$ are those curves $\gamma : I \longrightarrow \mathcal{M}$, for which the length functional

$$L[\gamma] \coloneqq \int_I \mathrm{d}t\, \mathscr{L}\Big(\dot\gamma(t), \dot\gamma(t), t\Big) = \int_I \mathrm{d}t \underbrace{\sqrt{g_{\gamma(t)}\Big(\gamma(t), \dot\gamma(t)\Big)}}_{=:\mathscr{L}} \tag{2.82}$$

is stationary. In a chart (U, θ) with $\gamma(I) \subset U$, this stationarity condition can be expressed as

$$0 \overset{!}{=} \frac{\delta L}{\delta \gamma^m(t)} = \frac{\partial \mathscr{L}}{\partial \gamma^m(t)} - \frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial \mathscr{L}}{\partial \dot{\gamma}^m(t)} \tag{2.83}$$

$$= \frac{1}{2} \left(\frac{\partial g_{ij}}{\partial \theta^m} - \frac{\partial g_{jm}}{\partial \theta^i} - \frac{\partial g_{mi}}{\partial \theta^j} \right) \dot{\gamma}^i(t)\, \dot{\gamma}^j(t) - g_{mj}\, \ddot{\gamma}^j(t) \tag{2.84}$$

and finally be brought to the familiar form given in literature by contraction with the inverse metric $-(g^{-1})^{am}$ on both sides, yielding

$$\ddot{\gamma}^a + \frac{1}{2} \left(g^{-1}\right)^{am} \left(\frac{\partial g_{jm}}{\partial \theta^i} + \frac{\partial g_{mi}}{\partial \theta^j} - \frac{\partial g_{ij}}{\partial \theta^m} \right) \dot{\gamma}^i\, \dot{\gamma}^j = 0. \tag{2.85}$$

Thus, any curve γ which satisfies this second order non-linear ordinary differential equation is a curve of extremal length between two points of the manifold. This means that a geodesic path represents either a minimum, a maximum or a saddle point of the length functional L. Depending on the geometry of the manifold, there may not exist both a longest and a shortest curve between two points as the length functional may not be bounded both from above and below.[21]

On Riemannian manifolds, one can alternatively define geodesics as curves which minimise the functional

$$E[\gamma] \coloneqq \int_I \mathrm{d}t \left\| \dot{\gamma}(t) \right\|^2 = \int_I \mathrm{d}t\, g_{\gamma(t)}\left(\dot{\gamma}(t), \dot{\gamma}(t)\right) = \int_I \mathrm{d}t \left(\mathscr{L}\left(\gamma(t), \dot{\gamma}(t)\right)\right)^2. \tag{2.86}$$

Since the instantaneous kinetic energy of a particle of mass m on a trajectory $\gamma(t)$ is classically given by $\frac{1}{2} m\, g_{\gamma(t)}\left(\dot{\gamma}(t), \dot{\gamma}(t)\right)$, one can interpret this functional as quantifying twice the energy of the curve. As one is only interested in minimisers of this functional anyway, this multiplicative factor can be omitted.

By the Cauchy–Schwarz inequality, one can relate the functionals $L[\gamma]$ and $E[\gamma]$ via

$$\left(L[\gamma]\right)^2 = \left[\int_I \mathrm{d}t\, 1 \cdot \left\| \dot{\gamma}(t) \right\| \right]^2 \leq \left[\int_I \mathrm{d}t\, 1^2 \right] \left[\int_I \mathrm{d}t \left\| \dot{\gamma}(t) \right\|^2 \right] = \mathrm{vol}(I)\, E[\gamma] \tag{2.87}$$

with the equality holding if the constant one-function $f(t) = 1$ and the norm $\|\dot{\gamma}\|$ are linearly dependent, i.e. if $\|\dot{\gamma}\|$ is constant for all $t \in I$ (see also [9, 54]).

2.6 Connections and Curvature

2.6.1 The Covariant Derivative

As developed so far, the formalism of differential geometry allows for the calculation of directional derivatives of functions using vectors. However, in order to define a chart-independent way of

[21] The existence of a solution to the geodesic ODE is guaranteed by the Picard–Lindelöf theorem, so any manifold equipped with a twice differentiable non-degenerate metric tensor field is guaranteed to have geodesics of some kind at least locally.

deriving tensor fields of arbitrary valence, a more powerful notion of derivative is necessary. There are two ways of achieving this: the covariant derivative and the Lie derivative, which generally lead to different results.

Whereas the Lie derivative does not need further structure to be introduced in addition to the a smooth atlas, the covariant derivative requires the choice of a so-called connection ∇. However, once a connection ∇ has been established, the covariant derivative is much more versatile and powerful than the Lie derivative. A direct comparison of the Lie derivative to the covariant derivative is postponed to section 2.9.2.

The action of the covariant derivative with respect to a vector field $X \in \Gamma(T\mathcal{M})$ on a general $\binom{p}{q}$-tensor field S is denoted by $\nabla_X S$ and again results in a $\binom{p}{q}$-tensor field. Its properties can be summarised as

(i) $\nabla_X f = Xf$

(ii) $\nabla_X (\alpha T + S) = \alpha \nabla_X T + \nabla_X S$

(iii) $\nabla_X (T \otimes S) = (\nabla_X T) \otimes S + T \otimes (\nabla_X S)$

(iv) $\nabla_{fX+Y} T = f \nabla_X T + \nabla_Y T$

where $T, S \in T^{(p,q)}\mathcal{M}$ are smooth tensor fields, $Y, Z \in \Gamma(T\mathcal{M})$ are smooth vector fields, $f \in C^\infty(\mathcal{M})$ and $\alpha \in \mathbb{R}$. In particular, the last property of the covariant derivative means that it is $C^\infty(\mathcal{M})$-linear in its lower slot, which distinguishes it from the Lie derivative. Condition (iii) directly implies

$$\nabla_X \Big(T(\omega, Y)\Big) = (\nabla_X T)(\omega, Y) + T(\nabla_X \omega, Y) + T(\omega, \nabla_X Y) \tag{2.88}$$

for the example of a $\binom{1}{1}$-tensor.

To work out precisely how much freedom remains in choosing a connection after enforcing the properties (i)–(iv), it is instructive to investigate the action of the covariant derivative in a chart (U, x)

$$\nabla_X Y = \nabla_{X^a \frac{\partial}{\partial x^a}} Y^b \frac{\partial}{\partial x^b} = X^a \left[\left(\nabla_{\frac{\partial}{\partial x^a}} Y^b \right) \frac{\partial}{\partial x^b} + Y^b \left(\nabla_{\frac{\partial}{\partial x^a}} \frac{\partial}{\partial x^b} \right) \right] \tag{2.89}$$

$$= X^a \left[\frac{\partial Y^b}{\partial x^a} \frac{\partial}{\partial x^b} + Y^b \left(\nabla_{\frac{\partial}{\partial x^a}} \frac{\partial}{\partial x^b} \right) \right]. \tag{2.90}$$

The last term of the previous equation is usually abbreviated as

$$\Gamma^i{}_{kj} := \mathrm{d}x^i \left(\nabla_{\frac{\partial}{\partial x^j}} \frac{\partial}{\partial x^k} \right) \tag{2.91}$$

and constitutes a family of $(\dim \mathcal{M})^3$ so-called connection coefficient functions which have to be provided explicitly in order to fully determine the connection. Oftentimes, the connection coefficient functions $\Gamma^i{}_{kj} \in C^\infty(\mathcal{M})$ are referred to as "Christoffel symbols of the second kind" in

the literature.[22]

Note that there are two different possible notations when it comes to the order of the lower indices of the Christoffel symbols. In the notation adopted here, the last lower index of the Christoffel symbol is determined by the index of the covariant derivative.

One can work out the following practical computational rules in a chart:

$$(\nabla_X Y)^a = X(Y^a) + \Gamma^a{}_{bc}\, Y^b\, X^c \tag{2.92}$$

$$(\nabla_X \omega)_b = X(\omega_b) - \Gamma^a{}_{bc}\, \omega_a\, X^c \tag{2.93}$$

$$(\nabla_X T)^a{}_b = X(T^a{}_b) + \Gamma^a{}_{dc}\, T^d{}_b\, X^c - \Gamma^d{}_{bc}\, T^a{}_d\, X^c \tag{2.94}$$

for vector fields $X, Y \in \Gamma(T\mathcal{M})$, a covector field $\omega \in \Gamma(T^*\mathcal{M})$ and a $\binom{1}{1}$-tensor field T and analogously for tensors of higher valence. Whenever it is clear from context that one only works in a single chart (U, x) such that there can be no confusion with a different chart basis, the abbreviation $\nabla_a \equiv \nabla_{\frac{\partial}{\partial x^a}}$ is frequently used.

Importantly, the sign difference for the Christoffel symbols when taking covariant derivatives of covector fields can be seen from

$$0 = \nabla_a\, \delta^b_c = \nabla_a \frac{\partial x^b}{\partial x^c} = \nabla_a \left[\mathrm{d}x^b \left(\frac{\partial}{\partial x^c} \right) \right] = \left(\nabla_a\, \mathrm{d}x^b \right) \left(\frac{\partial}{\partial x^c} \right) + \underbrace{\mathrm{d}x^b \left(\nabla_a \frac{\partial}{\partial x^c} \right)}_{= \Gamma^b{}_{ca}} \tag{2.95}$$

$$\Longrightarrow \left(\nabla_a\, \mathrm{d}x^b \right) \left(\frac{\partial}{\partial x^c} \right) = -\Gamma^b{}_{ca} \tag{2.96}$$

Since the Christoffel symbols fix the action of the covariant derivative on arbitrary tensor fields, this shows that the freedom one has in choosing a connection lies precisely in choosing the $(\dim \mathcal{M})^3$ coefficient functions. By plugging the transformation behaviour for tangent and cotangent space elements between two overlapping charts (U, x) and (V, y) into equation (2.91), one arrives at the following transformation behaviour in the overlap $U \cap V \subseteq \mathcal{M}$

$${}^{(x)}\Gamma^a{}_{bc} = {}^{(y)}\Gamma^r{}_{st} \frac{\partial x^a}{\partial y^r} \frac{\partial y^s}{\partial x^b} \frac{\partial y^t}{\partial x^c} + \frac{\partial x^a}{\partial y^k} \frac{\partial^2 y^k}{\partial x^b\, \partial x^c}. \tag{2.97}$$

Clearly, the Christoffel symbols do not transform like a $\binom{1}{2}$-tensor field in general, due to the fact that there exists a second term which depends non-linearly on the chart transition.[23] However, the difference of two Christoffel symbols does indeed transform like a $\binom{1}{2}$-tensor field, since the non-linear terms which are purely due to the chart transition map will always cancel each other.

Naturally this raises the question whether one can always find a transformation such that all the connection coefficient functions vanish. One can show that, due to the non-tensorial transformation

[22] The Christoffel symbols of the first kind are then given by contracting the first index with the metric $\tilde{\Gamma}_{ijk} := g_{im}\, \Gamma^m{}_{jk}$.

[23] If all components of a chart transition map $y \circ x^{-1}$ are linear then its second derivatives vanish. Therefore, under linear transformations, the effects of the non-linear term are not felt by the Christoffel symbols.

behaviour of the Christoffel symbols, such a transformation can indeed always be found. However, there is a caveat—it is only guaranteed that the Christoffel symbols vanish at a single point. In other words, there may not exist even an arbitrarily small neighbourhood around this point in which all Christoffel symbols vanish. Such coordinate charts in which the Christoffel symbols vanish (at least) at a point are referred to as normal coordinates.

As equation (2.105) shows, the Riemann tensor not only depends on the Christoffel symbols but also their derivatives, which means that normal coordinates do not allow one to transform away curvature. That is, scalar quantities built from the Riemann tensor such as the Ricci scalar remain invariant as expected.

2.6.2 Parallel Transport

Once a connection ∇ is established on a manifold, one defines a vector field X to be parallely transported along a smooth curve γ (or simply parallel to γ) if

$$\nabla_{\dot\gamma} X = 0. \tag{2.98}$$

A curve γ is said to be autoparallel (or autoparallely transported) if it satisfies

$$\nabla_{\dot\gamma} \dot\gamma = 0. \tag{2.99}$$

In physical terms, the left-hand side corresponds to the acceleration experienced by a particle travelling along γ. Therefore, the autoparallelity condition can be interpreted as the requirement that no unbalanced forces act on the particle meaning that autoparallels are precisely the curves corresponding to the uniform straight motion discussed by Newton's first axiom. In this sense, one may interpret Newton's first axiom as a measurement prescription for the connection on spacetime.

Since autoparallelity generalises the notion of straightness of a curve to curvilinear coordinates, autoparallels are often informally referred to as straight curves. In components, the autoparallelity condition can be expressed via

$$\left(\nabla_{\dot\gamma}\dot\gamma\right)^b = \dot\gamma^a \nabla_a \dot\gamma^b = \underbrace{\dot\gamma^a \frac{\partial \dot\gamma^b}{\partial x^a}}_{=\ddot\gamma^b} + \Gamma^b{}_{ma}\,\dot\gamma^m \dot\gamma^a = \ddot\gamma^b + \Gamma^b{}_{ma}\,\dot\gamma^m \dot\gamma^a \overset{!}{=} 0 \tag{2.100}$$

which is also known as the autoparallel equation.

By choosing the Levi-Civita connection one identifies the geodesic equation (2.85) with the autoparallel equation (2.100) which corresponds to requiring that geodesics, which are the stationary curves of the length functional, also coincide with the straight curves. Concretely, this identification amounts to inducing the connection coefficients Γ via the metric g according to

$$\Gamma^a{}_{bc} = \frac{1}{2}\left(g^{-1}\right)^{am}\left(\frac{\partial g_{bm}}{\partial x^c} + \frac{\partial g_{mc}}{\partial x^b} - \frac{\partial g_{bc}}{\partial x^m}\right) \tag{2.101}$$

in some chart (U, x). Moreover, it can be shown that the Levi-Civita connection is the unique connection which is both torsion-free (see section 2.6.3), as well as metric-compatible, i.e. $\nabla_X\, g = 0$ for all $X \in T\mathcal{M}$, which is occasionally also referred to as the fundamental theorem of Riemannian geometry.

2.6.3 Curvature and Torsion

The torsion T of a connection ∇ is the tensor field defined by

$$T(\omega, X, Y) \coloneqq \omega\Big(\nabla_X\, Y - \nabla_Y\, X - [X, Y]\Big) \quad \text{for} \quad \omega \in T^*\mathcal{M}, \quad X, Y \in T\mathcal{M}. \tag{2.102}$$

In addition to being C^∞-linear in each slot, it is also antisymmetric in the last slots, i.e. $T(\omega, X, Y) = -T(\omega, Y, X)$. In a chart (U, x), the torsion can be written as

$$T^i{}_{ab} = T\left(\mathrm{d}x^i, \frac{\partial}{\partial x^a}, \frac{\partial}{\partial x^b}\right) = \mathrm{d}x^i\left(\nabla_a \frac{\partial}{\partial x^b} - \nabla_b \frac{\partial}{\partial x^a}\right) = \Gamma^i{}_{ab} - \Gamma^i{}_{ba} \equiv 2\,\Gamma^i{}_{[ab]}. \tag{2.103}$$

where notation of square brackets around the indices implies the antisymmetrisation of the indices. Thus, a connection is torsion-free, i.e. the torsion tensor is zero everywhere, if and only if the Christoffel symbols are symmetric in their lower indices, i.e. $\Gamma^i{}_{ab} = \Gamma^i{}_{ba}$.

The Riemann tensor Riem which encodes the curvature information is defined as

$$\mathrm{Riem}(\omega, Z, X, Y) \coloneqq \omega\Big(\nabla_X\, \nabla_Y\, Z - \nabla_Y\, \nabla_X\, Z - \nabla_{[X,Y]}\, Z\Big). \tag{2.104}$$

for $\omega \in T^*\mathcal{M}$ and $X, Y, Z \in T\mathcal{M}$. Again, by evaluating the Riemann tensor on the basis and dual basis with respect to a chart, one immediately finds

$$\mathrm{Riem}^i{}_{jkl} = \mathrm{Riem}\left(\mathrm{d}x^i, \frac{\partial}{\partial x^j}, \frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^l}\right) = \frac{\partial}{\partial x^k}\,\Gamma^i{}_{jl} - \frac{\partial}{\partial x^l}\,\Gamma^i{}_{jk} + \Gamma^i{}_{ak}\,\Gamma^a{}_{jl} - \Gamma^i{}_{al}\,\Gamma^a{}_{jk}. \tag{2.105}$$

From its definition, one can immediately read off that for vector fields $X, Y \in T\mathcal{M}$ whose Lie bracket vanishes, i.e. $[X, Y] = 0$, the Riemann tensor quantifies the degree to which covariant derivatives fail to commute. In contrast to partial derivatives which always commute according to Schwarz's rule, the covariant derivative respects the shape and geometry of the underlying manifold.

In particular, one of the most important chart-independent measures of curvature is the so-called Ricci scalar R, which is defined by

$$R \coloneqq (g^{-1})^{ab}\,\mathrm{Riem}^c{}_{acb}. \tag{2.106}$$

For example, it can be shown that the Ricci scalar assumes a constant value of $R = 2$ on the sphere S^2 (given the canonical embedding into $\mathbb{R}^3$), which is why the sphere is said to be positively curved.

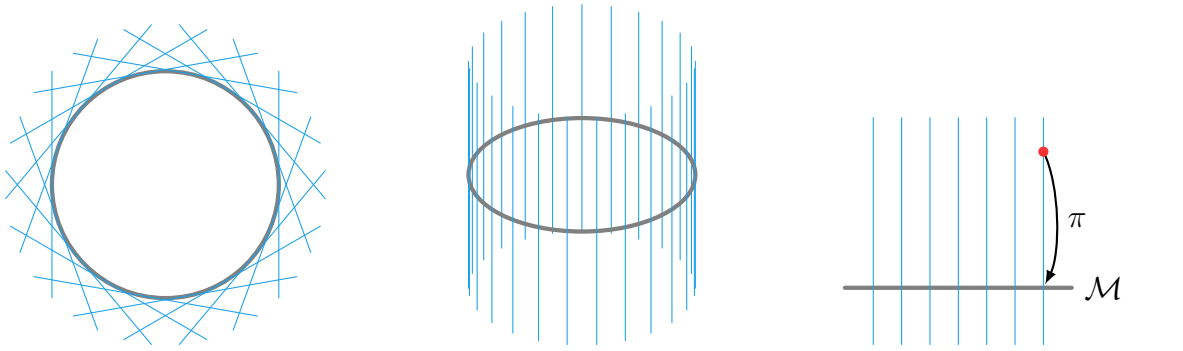
2.7 Fibre Bundles

A fibre bundle is a construct $E \xrightarrow{\pi} \mathcal{M}$ which consists of two smooth manifolds E and $\mathcal{M}$ and a smooth projection map π between them. Usually, E is referred to as the total space and $\mathcal{M}$ is called the base space. One can picture the total space as the base space $\mathcal{M}$ where at every point $p \in \mathcal{M}$ another space is attached which is called the fibre $F_p \coloneqq \operatorname{preim}_\pi(p) \equiv \left\{X \in E \;\middle|\; \pi(X) = p\right\} \subseteq E$.

Essentially, a fibre bundle provides a more general way of composing two spaces than a topological product. For example, while a cylinder can be written as a product of two topological spaces (i.e. the Cartesian product of their sets equipped with the product topology), namely $S^1 \times [0,1]$, it is not possible to express the Möbius strip as a product of two topological spaces due to its half-twist (see also figure 15).

Another natural example of a bundle is the tangent bundle $T\mathcal{M} \xrightarrow{\pi} \mathcal{M}$ where

$$T\mathcal{M} \coloneqq \dot{\bigcup_{p\in\mathcal{M}}} T_p\mathcal{M} \qquad \text{and} \qquad \pi(X) = p \qquad \forall X \in T_p\mathcal{M}. \tag{2.107}$$



(a) Tangent bundle of S^1 with the fibres drawn in the plane.

(b) Embedding of TS^1 in $\mathbb{R}^3$.

(c) Projection from an element of the fibre to its base point.

Figure 10: Visual representation of line bundles over the base space S^1 with the fibres drawn in blue. The right-hand side illustrates a projection map $\pi : E \longrightarrow \mathcal{M}$ which sends elements of the total space of a bundle to the unique point in the base space at which their respective fibre is attached.

Most importantly, fibre bundles allow for the definition of smooth sections

$$\sigma\colon \mathcal{M} \longrightarrow E \qquad \text{that satisfy} \qquad \pi \circ \sigma = \mathrm{id}_{\mathcal{M}} \tag{2.108}$$

that is, for any $p \in \mathcal{M}$, a section must always map to an element in the fibre over p. The notation

$$\Gamma(E) \coloneqq \{\sigma\colon \mathcal{M} \longrightarrow E \mid \pi \circ \sigma = \mathrm{id}_{\mathcal{M}},\ \sigma \text{ smooth}\} \tag{2.109}$$

is commonly used to denote the set of smooth sections of a bundle $E \xrightarrow{\pi} \mathcal{M}$. Smooth sections $\sigma \in \Gamma(T\mathcal{M})$ of the tangent bundle $T\mathcal{M}$ are usually referred to as vector fields. Importantly,

$\Gamma(T\mathcal{M})$ constitutes a so-called $C^\infty(\mathcal{M})$-module wherefore it is in particular also an infinite-dimensional $\mathbb{R}$-vector space.

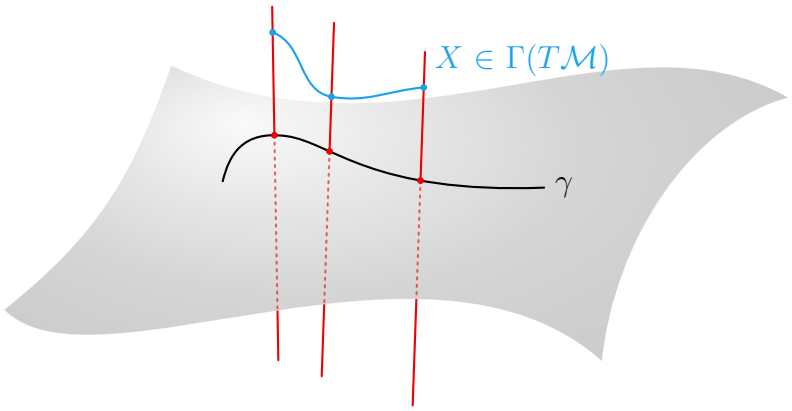


Figure 11: A smooth section $X \in \Gamma(T\mathcal{M})$ is visualised in blue over the tangent spaces (red) to a curve γ. This assignment of a particular fibre element to every point in a way that varies smoothly from point to point yields what one intuitively thinks of as a smooth vector field.

As already alluded to in section 2.4.1, it is possible to inherit an atlas from the base space $\mathcal{M}$ to the total space E along the projection map $\pi : E \longrightarrow \mathcal{M}$. In the particular case of the tangent bundle one can verify that the set of charts constructed by

$$\mathcal{A}_{T\mathcal{M}} \coloneqq \Big\{ (\mathrm{preim}_\pi(U), \xi) \;\Big|\; (U, x) \in \mathcal{A}_\mathcal{M} \Big\} \tag{2.110}$$

where the individual maps $\xi : \mathrm{preim}_\pi(U) \longrightarrow \mathbb{R}^{2\dim\mathcal{M}}$ are defined by

$$\xi(Z) \coloneqq \Big((x^1 \circ \pi)(Z), ..., (x^{\dim\mathcal{M}} \circ \pi)(Z), Z^1, ..., Z^{\dim\mathcal{M}} \Big) \mathrel{\widehat{=}} \Big((x \circ \pi)(Z), (\mathrm{d}x)(Z) \Big) \tag{2.111}$$

indeed constitutes a viable atlas on the total space of the tangent bundle. In particular, the projection map $\pi : (T\mathcal{M}, \mathcal{O}_{\text{initial}}, \mathcal{A}_{T\mathcal{M}}) \longrightarrow (\mathcal{M}, \mathcal{O}_\mathcal{M}, \mathcal{A}_\mathcal{M})$ is always rendered smooth by this construction. Likewise, it is possible to extend this definition to other fibre bundles.

2.8 Push-forward and Pull-back between Manifolds

For a smooth function $h : \mathcal{M} \longrightarrow \mathcal{N}$ between two smooth manifolds $\mathcal{M}$ and $\mathcal{N}$, one can define the associated push-forward map $h_* : T\mathcal{M} \longrightarrow T\mathcal{N}$ by its application on an arbitrary smooth function $f \in C^\infty(\mathcal{N})$

$$(h_*X)f \coloneqq X(f \circ h) \qquad \text{where} \qquad X \in T_{\pi_\mathcal{M}(X)}\mathcal{M}, \quad (h_*X) \in T_{h\left(\pi_\mathcal{M}(X)\right)}\mathcal{N}. \tag{2.112}$$

Since the vector that was pushed forward acts on the function f as the original vector field would have acted on the composition of the function f with the map h, this is a "natural" definition of how the tangent space elements of $\mathcal{M}$ and $\mathcal{N}$ should be related through h_*.

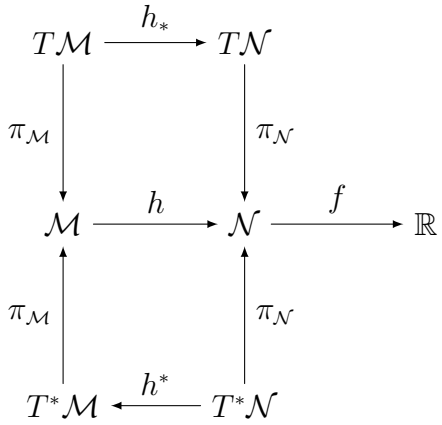


Figure 12: Illustration of how the tangent and cotangent spaces of two manifolds connected via a smooth map $h : \mathcal{M} \longrightarrow \mathcal{N}$ are related through the push-forward h_* and the pull-back h^*.

One can see from its definition that the push-forward map is $\mathbb{R}$-linear, i.e.

$$h_*(\alpha X) = \alpha(h_* X) \qquad \forall \alpha \in \mathbb{R}, \quad \forall X \in T\mathcal{M}. \tag{2.113}$$

For the case that $\mathcal{M} = \mathcal{N}$, the push-forward h_* is also called the derivative of h. Importantly, for a general map $h : \mathcal{M} \longrightarrow \mathcal{N}$, it is not possible to use the push-forward to obtain entire vector fields $Y \in \Gamma(T\mathcal{N})$ in the domain from vector fields $X \in \Gamma(T\mathcal{M})$ in the target. This is because the map h may not be surjective (i.e. it might not hit every point in the target) wherefore the resulting vector field $h_* X$ would not be defined everywhere on $\mathcal{N}$. On the other hand, if h is not injective, multiple vectors from $T\mathcal{M}$ would get mapped to the same element in $T\mathcal{N}$. Therefore, the push-forward of entire smooth vector fields is only well-defined under smooth bijective maps.

A particularly interesting application of the push-forward map is given if a lower-dimensional manifold $\mathcal{M}$ is embedded in a higher-dimensional manifold $\mathcal{N}$ since lower-dimensional tangent vectors are mapped to the higher-dimensional tangent vectors of the embedded manifold in the ambient space $\mathcal{N}$. An illustration of this is shown in figure 13.

Similarly, one can define the pull-back of covariant tensor fields $h^* : T^*\mathcal{N} \longrightarrow T^*\mathcal{M}$ by

$$(h^*\omega)(X) \coloneqq \omega(h_* X) \qquad \text{where} \qquad X \in T\mathcal{M}, \quad \omega \in \Gamma(T^*\mathcal{N}). \tag{2.114}$$

In particular, the pull-back of a $\binom{0}{2}$-tensor field such as the metric g is given by

$$(h^* g)(X, Y) = g(h_* X, h_* Y) \qquad \text{where} \qquad X, Y \in T\mathcal{M}, \quad g \in T^{(0,2)}\mathcal{N}. \tag{2.115}$$

In contrast to the push-forward, the pull-back of entire covector fields under smooth map is generally well-behaved. In particular, the special case of a pull-back $h^* f \in C^\infty(\mathcal{M})$ of a smooth

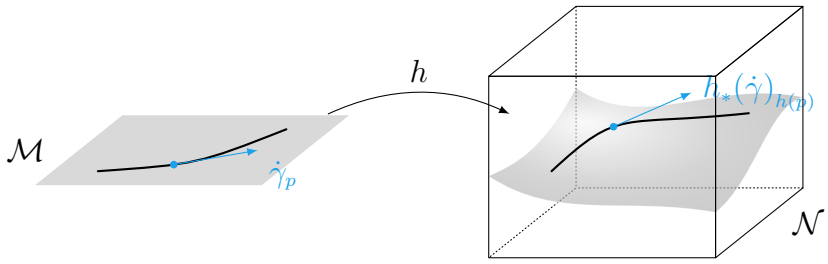


Figure 13: Illustration of the push-forward of tangent vectors along an embedding map $h : \mathcal{M} \longrightarrow \mathcal{N}$ in the case $\dim \mathcal{M} < \dim \mathcal{N}$. The view on the left-hand side is usually referred to as the intrinsic view of a manifold $\mathcal{M}$ whereas the right-hand side offers an extrinsic view of the manifold $\mathcal{M}$. The Nash embedding theorems guarantee that any Riemannian manifold can be isometrically embedded in some euclidean space, wherefore both ways of studying manifolds are ultimately equivalent.

function $f \in C^\infty(\mathcal{N})$ along $h : \mathcal{M} \longrightarrow \mathcal{N}$ is given by the composition of the two maps

$$h^* f = f \circ h. \tag{2.116}$$

Using the chart $(U \subseteq \mathcal{M}, x)$ on the initial manifold and $(V \subseteq \mathcal{N}, y)$ on the target manifold, one can compute the coefficient functions of the push-forward and the pull-back via

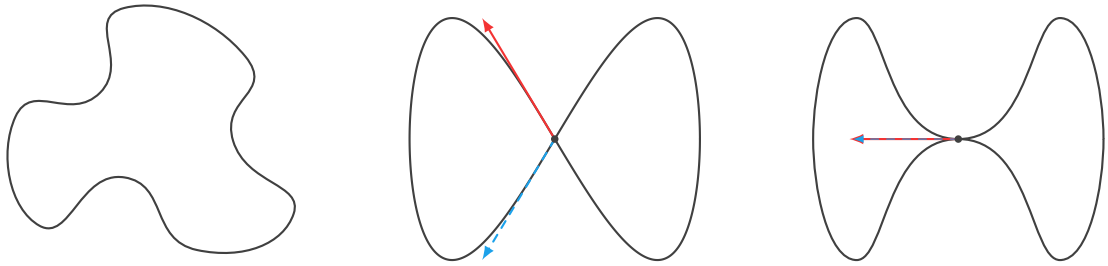
$$(h^*)^a{}_b = h^*(\mathrm{d}y^a)\left(\frac{\partial}{\partial x^b}\right) = \mathrm{d}y^a\left(h_*\left(\frac{\partial}{\partial x^b}\right)\right) = (h_*)^a{}_b \tag{2.117}$$

where careful attention must be paid to the fact that the indices a and b have different ranges and "live" in different manifolds. Concretely, $a = 1, ..., \dim \mathcal{N}$ and $b = 1, ..., \dim \mathcal{M}$. This shows that in charts, the component functions of the push-forward and pull-back coincide and thus, conveniently, one only has to compute them once. Whether a push-forward or a pull-back is performed is then encoded in the way in which indices are contracted. For example,

$$\left[h_*(X)\right]^a = (h_*)^a{}_b\, X^b \qquad \text{and} \qquad \left[h^*(\omega)\right]_b = (h^*)^a{}_b\, \omega_a = (h_*)^a{}_b\, \omega_a \tag{2.118}$$

where $X \in \Gamma(T\mathcal{M})$ and $\omega \in \Gamma(T^*\mathcal{N})$.

As demonstrated by figure 14, the difference between immersions and embeddings is whether they feature self-intersections. Another well-known example of an immersion is given by the Klein bottle, which cannot be mapped into $\mathbb{R}^3$ without self-intersection.



(a) Embedding **(b)** Immersion **(c)** Neither

Figure 14: Illustration of smooth maps from the circle S^1 into the plane $\mathbb{R}^2$ which highlight the difference between embeddings, immersions and maps which qualify as neither. Apart from being smooth, embedding maps are required to be injective, wherefore distinct points in the domain may be not mapped to the same point in the image. In contrast, an immersion f is not required to be injective itself, but its push-forward f_* is. Therefore, although the image $f(S^1)$ of an immersion f may intersect itself, no two tangent vectors of the domain are allowed to be mapped to the same vector in the target.

2.9 Lie Groups

A group $(G, \diamond)$ is a set G together with an operation $\diamond : G \times G \longrightarrow G$ which satisfies

$$\text{(Associativity)} \qquad \forall g_1, g_2, g_3 \in G : \qquad g_1 \diamond (g_2 \diamond g_3) = (g_1 \diamond g_2) \diamond g_3 \tag{2.119}$$

$$\text{(Identity element)} \qquad \exists e \in G : \forall g \in G : \qquad e \diamond g = g \diamond e = g \tag{2.120}$$

$$\text{(Inverse element)} \qquad \forall g \in G : \exists g^{-1} \in G : \qquad g \diamond g^{-1} = g^{-1} \diamond g = e. \tag{2.121}$$

If, additionally, the group operation $\diamond$ is commutative, i.e.

$$\forall g_1, g_2 \in G : \qquad g_1 \diamond g_2 = g_2 \diamond g_1, \tag{2.122}$$

then the group is said to be abelian.

A Lie group is a group $(G, \diamond)$ where the underlying set G is a smooth manifold and also the group operation $\diamond$ is smooth in the sense that the two maps μ, i defined by

$$\mu : G \times G \longrightarrow G \qquad \mu(g_1, g_2) \coloneqq g_1 \diamond g_2 \tag{2.123}$$

$$i : G \longrightarrow G \qquad i(g) \coloneqq g^{-1} \tag{2.124}$$

are both smooth. An important example of a Lie group is given by the so-called general linear group $\mathrm{GL}(n, \mathbb{R})$, which is the group of all invertible real $n \times n$ matrices where the group operation is given by the usual matrix multiplication. Moreover, it contains many other interesting subgroups such as the special orthogonal group $\mathrm{SO}(n)$, which is given by the set of all $n \times n$ orthogonal matrices with unit determinant and is generally referred to as the rotation group. A simpler example of a group is given by $(\mathbb{R}, +)$ which is the set of real numbers with addition as the group operation. Due to the fact that Lie groups are intimately connected to the notion of continuous symmetries, they play a key role in fundamental physics.

2.9.1 Lie Algebras

A Lie algebra $\mathfrak{L} = (V, [\,\cdot\,,\,\cdot\,])$ is a vector space V, equipped with a bracket[24] operation $[\,\cdot\,,\,\cdot\,] : V \times V \longrightarrow V$ which satisfies for all elements $X, Y, Z \in V$

$$\text{(Linearity)} \qquad [\alpha X + Y, Z] = \alpha[X, Z] + [Y, Z] \tag{2.125}$$

$$\text{(Antisymmetry)} \qquad [X, Y] = -[Y, X] \tag{2.126}$$

$$\text{(Jacobi-identity)} \qquad \big[X, [Y, Z]\big] + \big[Y, [Z, X]\big] + \big[Z, [X, Y]\big] = 0. \tag{2.127}$$

A tuple $(W, [\,\cdot\,,\,\cdot\,])$ forms a Lie subalgebra of $(V, [\,\cdot\,,\,\cdot\,])$ if W is a vector subspace of V and is closed with respect to the shared Lie bracket, i.e. if the Lie bracket of any two elements in W is again an element of W.

An important example is given by the infinite-dimensional Lie algebra of smooth vector fields $(\Gamma(T\mathcal{M}), [\,\cdot\,,\,\cdot\,])$ on a manifold[25] where the corresponding Lie bracket is defined via

$$[X, Y]f \coloneqq X(Yf) - Y(Xf) \qquad \forall f \in C^\infty(\mathcal{M}). \tag{2.128}$$

For a finite-dimensional Lie algebra $\mathfrak{L}$, one can find a basis $X_1, \ldots, X_n \in \mathfrak{L}$ with respect to which the action of the Lie bracket can be captured in the form of so-called structure constants $C^k{}_{ij} \in \mathbb{R}$ defined by

$$[X_i, X_j] = C^k{}_{ij}\, X_k. \tag{2.129}$$

Using the structure constants of a finite-dimensional Lie algebra $\mathfrak{L}$, one can define a symmetric bilinear form[26] known as the Killing form B via

$$B_{ij} = C_{im}{}^n\, C_{jn}{}^m. \tag{2.130}$$

It can be shown that the Killing form B is non-degenerate if and only if the Lie algebra $\mathfrak{L}$ can be decomposed into a direct sum of simple Lie algebras.

Crucially, every Lie group G gives rise to a unique Lie algebra, which is defined as the set of left-invariant vector fields on G, that is, vector fields $X \in \Gamma(TG)$ which satisfy

$$\forall g \in G : \qquad (l_g)_* X = X \tag{2.131}$$

where l_g is the left-action[27] produced by the group element $g \in G$ defined by $l_g(h) \coloneqq g \diamond h$. The Lie bracket on the set of left-invariant vector fields is then chosen as the usual Lie bracket of smooth vector fields $\Gamma(TG)$. Furthermore, the set of left-invariant vector fields can be identified with the

[24] Depending on the context, this bracket is often referred to as the commutator bracket.

[25] More specifically, one can show that $\Gamma(T\mathcal{M})$ is a $C^\infty(\mathcal{M})$-module, which implies that it is in particular also an $\mathbb{R}$-vector space.

[26] In this context, the term bilinear form is not connected to the notion of differential forms.

[27] Since one can check that this left-action constitutes a diffeomorphism on the Lie group G, the push-forward with respect to the left-action map is well-defined for entire vector fields.

tangent space at the identity of the group, i.e. $\mathrm{Lie}(G) \cong_{\text{Lie alg.}} T_eG$. However, while every Lie group has a unique corresponding Lie algebra, the converse is not true—different (non-isomorphic) Lie groups can have the same Lie algebra. It is common to denote the Lie algebra corresponding to a Lie group G using the lowercase $\mathfrak{g}$.

2.9.2 Lie Derivatives and Symmetry

The flow of a nowhere vanishing smooth vector field $X \in \Gamma(T\mathcal{M})$ is a one-parameter family

$$h^X : \mathbb{R} \times \mathcal{M} \longrightarrow \mathcal{M}, \qquad (t,p) \longmapsto h^X(t,p) \equiv h^X_t(p) \coloneqq \gamma_{(p)}(t) \tag{2.132}$$

where $\gamma_{(p)}(t)$ is the unique integral curve (see sections 4.3 and 4.3.3) of X with $\gamma(0) = p$. Thus, for a fixed value of t, $h^X_t : \mathcal{M} \longrightarrow \mathcal{M}$ is a smooth map. The fact that this is a group action can be seen from

$$h^X_s \diamond h^X_t = h^X_{s+t} \qquad \text{and} \qquad \forall t \in \mathbb{R} : \exists(-t) \in \mathbb{R} : \qquad h^X_{-t} \diamond h^X_t = h^X_0 = \mathrm{id}_{\mathcal{M}} \,. \tag{2.133}$$

Flows of vector fields offer an alternative way of describing the integral curves generated by said vector field through Lie group actions. It is then natural to define a covariant tensor field g to be symmetric[28] with respect to a vector field X, if g does not change along the flow of X, which can be encoded as

$$(h^X_t)^* g \overset{!}{=} g \qquad \Longleftrightarrow \qquad (h^X_t)^* g - g = 0. \tag{2.134}$$

By dividing this definition through the parameter t and considering the limit as $t \longrightarrow 0$, one arrives at the so-called Lie derivative of g of the which is defined as

$$\mathcal{L}_X \, g \coloneqq \lim_{t \to 0} \frac{(h^X_t)^* g - g}{t}. \tag{2.135}$$

More generally, the Lie derivative $\mathcal{L}_X$ with respect to a vector field $X \in \Gamma(T\mathcal{M})$ on a smooth manifold $(\mathcal{M}, \mathcal{O}, \mathcal{A})$ is a map $\mathcal{L}_X : T^{(p,q)}\mathcal{M} \longrightarrow T^{(p,q)}\mathcal{M}$ which associates $\binom{p}{q}$-tensor fields to $\binom{p}{q}$-tensor fields such that it satisfies $\forall X, Y \in \Gamma(T\mathcal{M})$, $\forall Q, S \in T^{(p,q)}\mathcal{M}$, $\forall f \in C^\infty(\mathcal{M})$ and $R \in T^{(1,1)}\mathcal{M}$:

(i) $\mathcal{L}_X \, f = Xf$

(ii) $\mathcal{L}_X \, Y = [X, Y]$

(iii) $\mathcal{L}_X \, (Q + S) = \mathcal{L}_X \, Q + \mathcal{L}_X \, S$

(iv) $\mathcal{L}_X \Big(R(\omega, Y) \Big) = (\mathcal{L}_X \, R)(\omega, Y) + R(\mathcal{L}_X \, \omega, Y) + R(\omega, \mathcal{L}_X \, Y)$

(v) $\mathcal{L}_{X+Y} \, S = \mathcal{L}_X \, S + \mathcal{L}_Y \, S$.

Most importantly, the covariant derivative is $C^\infty(\mathcal{M})$-linear in its lower slot, while the Lie derivative is only $\mathbb{R}$-linear in the lower slot. When acting on a $\binom{0}{0}$-tensor field, that is, a function, the Lie

[28] Importantly, symmetries with respect to groups are in no way connected to (anti-)symmetries of a tensor in its indices.

derivative exactly coincides with the covariant and the exterior derivative by construction.

A $\binom{p}{q}$-tensor field S is said to be symmetric with respect to a Lie algebra $(\mathfrak{L}, [\,\cdot\,,\,\cdot\,])$ if

$$\forall X \in \mathfrak{L}: \qquad \mathcal{L}_X S = 0 \tag{2.136}$$

that is, the Lie derivative of S vanishes with respect to all elements of the Lie algebra $\mathfrak{L}$. If a basis of $\mathfrak{L}$ is known, it generally suffices to check this property for the basis elements of $\mathfrak{L}$.

Using the properties of the Lie derivative listed above, one obtains the following practical computational rules in a chart (U, x)

$$(\mathcal{L}_X Y)^a = X(Y^a) - Y^b \frac{\partial X^a}{\partial x^b} \tag{2.137}$$

$$(\mathcal{L}_X \omega)_a = X(\omega_a) + \omega_b \frac{\partial X^b}{\partial x^a} \tag{2.138}$$

$$(\mathcal{L}_X S)^a{}_b = X(S^a{}_b) - S^m{}_b \frac{\partial X^a}{\partial x^m} + S^a{}_m \frac{\partial X^m}{\partial x^b} \tag{2.139}$$

where $X, Y \in \Gamma(T\mathcal{M})$, $\omega \in \Gamma(T^*\mathcal{M})$ and $S \in T^{(1,1)}\mathcal{M}$. Similarly, these rules apply for tensors of other valence. A direct comparison with equations (2.92) to (2.94) reveals that the sign dependence of terms on whether the derived quantity is covariant or contravariant is exactly reversed for the Lie derivative.

As already alluded to in section 2.6.1, while it does not require the introduction of further structure beyond a smooth atlas, the Lie derivative is not as versatile as the covariant derivative induced by a connection. For example, the covariant derivative of a $\binom{1}{1}$-tensor field S can be calculated with respect to a vector $X_p \in T_p\mathcal{M}$ at a single point $p \in \mathcal{M}$ by

$$(\nabla_{X_p} S)^b{}_c = X^a|_p \left[\frac{\partial S^b{}_c}{\partial x^a} + \Gamma^b{}_{ja} S^j{}_c - \Gamma^j{}_{ca} S^b{}_j\right]_p \tag{2.140}$$

On the other hand, equation (2.139) clearly shows that the Lie derivative $\mathcal{L}_X S$ depends on derivatives of the vector field X. Therefore, knowledge of the component functions of X is required at least in a finite (non-empty) neighbourhood around the point $p \in \mathcal{M}$. As a result, it is not possible to calculate the Lie derivative of tensors (whose rank is non-zero) with respect to a vector at a single point. By extension, it is also impossible to calculate the Lie derivative of a tensor (whose rank is non-zero) along a curve γ, i.e. it is impossible to evaluate the expression $\mathcal{L}_\gamma S$.

Thus, one can conclude from this comparison with the covariant derivative that the Lie derivative compensates for the lack of further structure by requiring not only the components of a vector at every point, but also of the first derivatives of its components.

2.10 Integration on Manifolds

In order to lift the notion of integration to the manifold level in such a way that its definition does not rely on any choice of specific coordinates, a few structural refinements of a smooth manifold

$(\mathcal{M}, \mathcal{O}, \mathcal{A})$ must be made:

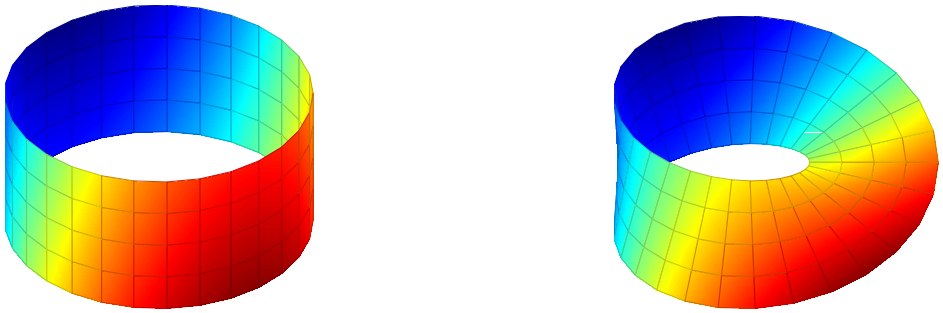
First, one must choose an orientation by restricting the atlas $\mathcal{A}$ in such a way that any chart transition preserves this orientation. Second, a volume form Ω must be chosen, which, as the name suggests, quantifies the volume contained in a parallelotope spanned by $(\dim \mathcal{M})$-many vectors. Third, a so-called partition of unity which smoothly stitches together overlapping charts must be specified for all regions of chart overlap contained in the integration domain.

The need to fix the orientation ultimately arises because it must be ensured that integrations which have been split up over multiple charts do not inadvertently cancel out when their contributions should, in fact, add (and vice versa).

An atlas $\mathcal{A}^{\uparrow} \subseteq \mathcal{A}$ is said to be positively oriented if for any overlapping pair of charts the determinant of the Jacobian is positive, i.e.

$$\forall (U, x), (V, y) \in \mathcal{A}^{\uparrow} : \forall p \in U \cap V : \qquad \det\left(\frac{\partial y}{\partial x}\bigg|_p \right) > 0. \tag{2.141}$$

Conversely, for a negatively oriented atlas $\mathcal{A}^{\downarrow}$ the above inequality for the Jacobian determinant is reversed. It is important to note that not all smooth manifolds admit an oriented atlas: a famous example of a smooth manifold on which no consistent orientation can be chosen is given by the Möbius strip (see figure 15).



(a) Cylinder. **(b)** Möbius strip.

Figure 15: Illustration of a cylinder and Möbius strip embedded in $\mathbb{R}^3$. Whereas the cylinder is two-sided and two-edged, the Möbius strip is a one-sided surface and only has a single edge as a result of its half twist. Thus, it is impossible to find a consistently oriented atlas covering the whole Möbius strip.

Second, one must choose a nowhere-vanishing volume form Ω, i.e. a smooth and totally antisymmetric $\binom{0}{\dim \mathcal{M}}$-tensor field. In particular, the choice of such a volume form is unique up to a (nowhere-vanishing) smooth function. If in addition to the oriented smooth structure, the manifold $\mathcal{M}$ is also equipped with a Riemannian metric g, then the canonical choice for this volume form Ω_g in a chart $(U, x) \in \mathcal{A}^{\uparrow}$ is given by

$$\Omega_g \coloneqq \underbrace{\sqrt{\det(g)}}_{=:\omega_g} \, \mathrm{d}x^1 \wedge \ldots \wedge \mathrm{d}x^n = \sqrt{\det(g)} \, \varepsilon_{i_1 \ldots i_n} \, \mathrm{d}x^{i_1} \ldots \mathrm{d}x^{i_n} \tag{2.142}$$

where $n = \dim \mathcal{M}$ and ε is the Levi-Civita symbol. Therefore, the metric-induced volume form Ω_g already fixes the choice of a non-vanishing smooth function as $\omega_g = \sqrt{\det(g)}$. It is straightforward to verify that Ω_g is indeed chart-independent in that the factors produced by ω_g and $\mathrm{d}x^1 \wedge \ldots \wedge \mathrm{d}x^n$ under a change of chart cancel out up to a factor of $\mathrm{sgn}[\det(\partial y/\partial x)]$ which is also where the need to choose a consistent orientation becomes apparent.

Third, a partition of unity is a finite[29] set R of continuous / smooth functions $\rho_i : \mathcal{M} \longrightarrow [0, 1]$ which collectively satisfy

$$\sum_{i \in I} \rho_i(z) = 1 \tag{2.143}$$

at every point $z \in \mathcal{M}$ with I an index set of R. In particular, the idea is to associate one such function ρ_i to every chart (V_i, y_i) in a finite cover of the integration domain U in the sense that $\rho_i(z) = 0$ if $z \notin V_i$. Thereby, it is ensured that overlap regions of charts do not result in over-counting of the evaluated function. This formalises the idea of smoothly blending together the constituent integrals which are over separate charts.

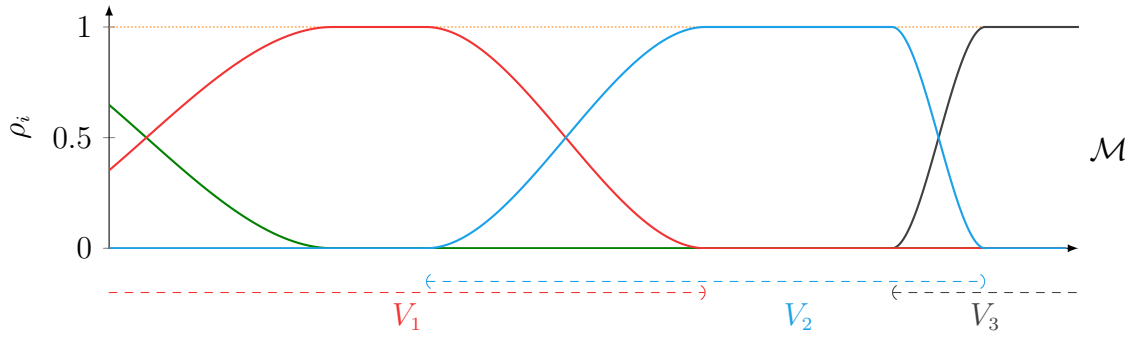


Figure 16: Illustration of the smooth interpolation of overlapping charts through a partition of unity.

A well-known result from topology guarantees that for topological spaces which are Hausdorff, a partition of unity exists if and only if they are paracompact (see section 2.1). Both of these criteria are fulfilled in typical applications since only Riemannian manifolds that can be covered using a finite number of charts are considered here.

The integral of a function $f : \mathcal{M} \longrightarrow \mathbb{R}$ over some open (i.e. measurable) domain $U \subseteq \mathcal{M}$ is then defined as

$$\int_U f \coloneqq \sum_{i \in I} \int_{V_i} \rho_i \cdot \mu_U \cdot f \qquad \text{where} \qquad \mu_U(z) \coloneqq \begin{cases} 1 & z \in U \\ 0 & z \in \mathcal{M} \setminus U \end{cases} \tag{2.144}$$

is an indicator function and I an index set of the partition of unity R. Each integration over a

[29] Technically, the partition of unity may contain infinitely many functions ρ_i and it is only necessary to ensure that there exists a neighbourhood around any point in which only a finite number of ρ_i are non-zero to avoid convergence issues of $\sum_i \rho_i$.

single chart (V, y) can be reduced to standard Lebesgue integrals by

$$\int_V \tilde{f} = \int_{y(V)\subseteq\mathbb{R}^n} (y^{-1})^*(\Omega \cdot \tilde{f}) = \int_{y(V)\subseteq\mathbb{R}^n} \mathrm{d}^n\alpha\, (\omega_g \circ y^{-1})(\alpha) \cdot (\tilde{f} \circ y^{-1})(\alpha) \tag{2.145}$$

which can then be executed using standard techniques. As before, ω_g denotes the so-called geometric density function induced by the metric g on $\mathcal{M}$ and $n = \dim \mathcal{M}$.

2.11 Probability Theory

2.11.1 Kolmogorov Axioms of Probability Theory

The modern axiomatic formulation of probability theory was conceived by Andrey Kolmogorov in the 1930s. Given a measure space $(\mathcal{M}, \mathcal{B}, \mathbb{P})$ with $\mathcal{B}$ the Borel σ–algebra, Kolmogorov's three axioms can be summarised as follows:

(Positivity) $$\mathbb{P}(U) \geq 0 \qquad \forall U \in \mathcal{B} \tag{2.146}$$

(Normalisation) $$\mathbb{P}(\mathcal{M}) = 1 \tag{2.147}$$

(σ–additivity) $$\mathbb{P}\Big(\dot{\bigcup_{j\in J}} U_j\Big) = \sum_{j\in J} \mathbb{P}(U_j) \qquad \text{with } J \text{ a countable index set.} \tag{2.148}$$

If the measure space $(\mathcal{M}, \mathcal{B}, \mathbb{P})$ satisfies the Kolmogorov axioms, it is commonly referred to as a probability space. Moreover, it is straightforward to show that if the Kolmogorov axioms are fulfilled, then the probability measure $\mathbb{P}$ also satisfies for all $A, B \in \mathcal{B}$

(Empty set) $$\mathbb{P}(\varnothing) = 0 \tag{2.149}$$

(Complement) $$\mathbb{P}(A^c) = 1 - \mathbb{P}(A) \tag{2.150}$$

(Monotonicity) $$A \subseteq B \quad \Longrightarrow \quad \mathbb{P}(A) \leq \mathbb{P}(B). \tag{2.151}$$

Although a measure-theoretic formulation is both more general and rigorous, it is also considerably more technical. Since the added generality is of no benefit in the context of this thesis, the usual explicit notation where probability density functions p are considered directly is adopted from here on out. For literature on probability and information theory which does emphasise a measure-theoretic approach, see for example [8, 26].

2.11.2 Elementary Objects of Probability Theory

A family of probability distributions is a set where each distribution $p(x;\theta)$ in the set is usually characterised by a parameter configuration $\theta = (\theta^1, \ldots, \theta^n)$ which varies continuously over some open set in $\mathbb{R}^n$ (see [49]). Thus, any particular member of a family of probability distributions uniquely corresponds a choice of parameters $\theta \in \mathcal{M}$.

An important example is given by the family of normal distributions, which is generally denoted

by

$$N = \Big\{ p(x;\mu,\sigma) \;\Big|\; \mu,\sigma \in \mathbb{R},\, \sigma > 0 \Big\} \qquad \text{and} \qquad \Big[N(\mu,\sigma)\Big](x) \mathrel{\widehat{=}} p(x;\mu,\sigma). \tag{2.152}$$

Accordingly, each member of the normal family has a probability density function of the form

$$p(x;\theta) \equiv p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}\, \exp\!\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \tag{2.153}$$

where the parameter θ is the pair (μ,σ), i.e. the mean and standard deviation. In the case of the normal family, a choice of parameters (μ,σ) can be interpreted as the Cartesian coordinates of a point in the upper half plane of $\mathbb{R}^2$ (see [49]).

The expectation value of a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ with respect to a probability distribution $p(x)$ is denoted by

$$\Big\langle f(x) \Big\rangle_x \equiv \mathbb{E}_x\Big(f(x)\Big) \coloneqq \int_{\mathbb{R}^n} \mathrm{d}^n x\; p(x)\, f(x) \tag{2.154}$$

where the combination $(p \cdot f)$ must be an integrable function for the expectation value to exist.[30] The subscript x in $\langle\,\cdot\,\rangle_x$ and $\mathbb{E}_x(\,\cdot\,)$ indicates the integration variable but is often omitted whenever one can unambiguously infer it from the context.[31]

The variance of a function $f(x)$ with respect to the probability density is then defined as

$$\mathrm{Var}(f) \coloneqq \mathbb{E}_x\Big(\big(f(x) - \mathbb{E}(f)\big)^2\Big) \equiv \Big\langle \big(f(x) - \langle f \rangle\big)^2 \Big\rangle_x = \mathbb{E}\big(f^2\big) - \mathbb{E}(f)^2. \tag{2.155}$$

From the definition, one can check that the variance satisfies for all $\alpha, \beta \in \mathbb{R}$

$$\mathrm{Var}_x\Big(\alpha\, f(x) + \beta\Big) = \alpha^2\, \mathrm{Var}(f). \tag{2.156}$$

The n-th moment m_n of a probability distribution is given by $m_n \coloneqq \mathbb{E}(x^n)$. Of special interest are often the first four moments of a distribution. The moment-generating function $M_x(t)$ of a probability distribution is defined as

$$M_x(t) \coloneqq \mathbb{E}\big(e^{tx}\big) = \mathbb{E}\left(\sum_{n=0}^{\infty} \frac{t^n\, x^n}{n!}\right) = \sum_{n=0}^{\infty} \frac{t^n}{n!}\, \mathbb{E}(x^n) = 1 + t\,\mathbb{E}(x) + \frac{t^2\,\mathbb{E}(x^2)}{2!} + \ldots \tag{2.157}$$

where $t \in \mathbb{R}$.[32] The moment-generating function gets its name from the fact that the n-th moment of a probability distribution is recovered by

$$\frac{\partial^n M_x(t)}{\partial t^n}\bigg|_{t=0} = \left[\mathbb{E}(x^n) + t\,\mathbb{E}\big(x^{n+1}\big) + \frac{t^2}{2!}\,\mathbb{E}\big(x^{n+2}\big) + \ldots \right]_{t=0} = \mathbb{E}(x^n) = m_n \tag{2.158}$$

[30] Since integrals are instances of linear functionals, taking the expectation value is also a linear operation.

[31] Another common notation is to indicate the probability distribution p with respect to which the expectation value is taken in the subscript as $\langle\,\cdot\,\rangle_p$ or even more explicitly as $\langle\,\cdot\,\rangle_{p(x)}$.

[32] The moment-generating function does not always exist, since the integral involved in the expectation value may diverge for some (or all) orders.

Instead of using moments, it is typically more convenient to work with the cumulants κ_n of a probability distribution

$$\kappa_n := \left.\frac{\partial^n \ln\Big(M_x(t)\Big)}{\partial t^n}\right|_{t=0} \tag{2.159}$$

which in turn can be extracted from the cumulant-generating function

$$K_x(t) := \ln\Big(M_x(t)\Big) = \ln\Big(\mathbb{E}\Big(e^{tx}\Big)\Big). \tag{2.160}$$

Similar to the moment-generating function $M_x(t)$, one defines the characteristic function as

$$\varphi_x(t) := \mathbb{E}\Big(e^{\mathrm{i}tx}\Big) = M_x(\mathrm{i}t) \tag{2.161}$$

where i denotes the imaginary unit. The reason why one is interested in characteristic functions is that they are dual to the probability density function of the same underlying random variable x with respect to the Fourier transform. That is,

$$p(x) = (\mathcal{F}\varphi)(x) \qquad \text{and} \qquad \varphi_x(t) = \Big(\mathcal{F}^{-1}\, p\Big)(t) \tag{2.162}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier transform and its inverse transform respectively and are defined as[33]

$$\tilde{p}(t) = (\mathcal{F}\, p)(t) := \int \mathrm{d}x\, e^{-\mathrm{i}tx}\, p(x) \quad \text{and} \quad \Big(\mathcal{F}^{-1}\, \tilde{p}\Big)(x) := \int \frac{\mathrm{d}t}{2\pi}\, e^{\mathrm{i}tx}\, \tilde{p}(t). \tag{2.163}$$

The main reason why it is convenient to calculate the above generating functions is that it transforms the integrations which are associated with taking expectations and are necessary to calculate each moment or cumulant individually into differentiations which are significantly easier to carry out in practice. Thus, only a single integration must be carried out to find the generating function and from then on only differentiation is necessary to find moments and cumulants of any order.

The cumulative distribution function $F(x)$ corresponding to a univariate probability density function $p(x)$ is defined as

$$F(x) := \int_{-\infty}^{x} \mathrm{d}z\, p(z) \tag{2.164}$$

from which it is clear that it is always a monotone increasing function since probability densities are non-negative. Especially the inverse of cumulative distributions F play a vital role in hypothesis testing and are defined as

$$F^{-1}(u) = \inf\Big\{x \;\Big|\; u \leq F(x)\Big\}. \tag{2.165}$$

Typically, the inverse of the cumulative distribution is referred to as the quantile function associated with p.

[33] The convention of asymmetric prefactors for the Fourier transform and its inverse are chosen here, which is the standard convention throughout most physics literature.

Given that integrals of univariate probability distributions[34] p which are defined on the whole real line $\mathbb{R}$ can be expressed using their cumulative distribution F as

$$\int_{-c}^{c} \mathrm{d}x\, p(x) = F(c) - F(-c). \tag{2.166}$$

For probability densities which are symmetric around $x = 0$ and thus satisfy $p(x) = p(-x)$, one has

$$\int_{-c}^{c} \mathrm{d}x\, p(x) = 2\int_{0}^{c} \mathrm{d}x\, p(x) = 2\Big[F(c) - \underbrace{F(0)}_{1/2}\Big] = 2F(c) - 1 \tag{2.167}$$

which allows one to express the cumulative normal normal distribution with mean $\mu = 0$ as

$$F(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right) \tag{2.168}$$

Here, erf denotes the Gaussian error function which frequently comes up in the analysis normal distributions and is defined by

$$\operatorname{erf}(x) \coloneqq \pi^{-\frac{1}{2}} \int_{-x}^{x} \mathrm{d}t\, e^{-t^2} = \frac{2}{\sqrt{\pi}} \int_{0}^{x} \mathrm{d}t\, e^{-t^2}. \tag{2.169}$$

For example, the volume of the $n\sigma$ confidence interval of a normal distribution can be expressed as

$$\int_{-n\sigma}^{n\sigma} \mathrm{d}x\, p(x; 0, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-n\sigma}^{n\sigma} \mathrm{d}x \exp\left(-\frac{x^2}{2\sigma^2}\right) = \operatorname{erf}\left(\frac{n}{\sqrt{2}}\right). \tag{2.170}$$

The confidence level $q \in [0, 1]$ of a confidence region is a measure of the probability volume it contains. Conversely, the significance level $\alpha = 1 - q$ quantifies the probability volume outside a confidence region. It is common to specify confidence levels in terms of the standard deviation σ of a normal distribution. That is, instead of referring to a confidence region as being approximately of confidence level 0.6827, one uses the abbreviation of a 1σ confidence level. A convenient way to convert between these two ways of representing the confidence level is afforded by the Gaussian error function

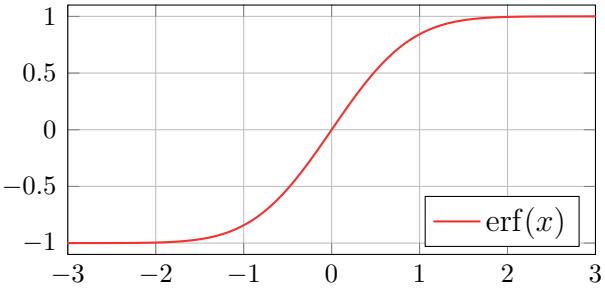
$$\operatorname{erf}\left(\frac{n}{\sqrt{2}}\right) = q = 1 - \alpha \tag{2.171}$$

where n is the number of the confidence interval in units of the standard deviation σ. A table of frequently-used numerical values of the Gaussian error function as well as a plot can be found in figure 17. Additionally, a plot of the confidence intervals of a standard normal distribution is depicted in figure 18.

[34] In the multivariate case, the joint cumulative distribution can be defined as $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$ and analogously for higher dimensions.

Confidence Level n	$\mathrm{erf}(n/\sqrt{2})$
1	0.6826895
2	0.9544997
3	0.9973002
4	0.9999367
5	0.9999994

(a) Useful values of the error function.



(b) Plot of the error function.

Figure 17: The left-hand side table summarises the probability volumes contained in the confidence intervals of levels 1σ to 5σ for a univariate normal distribution. The right-hand side shows a plot of the Gaussian error function.

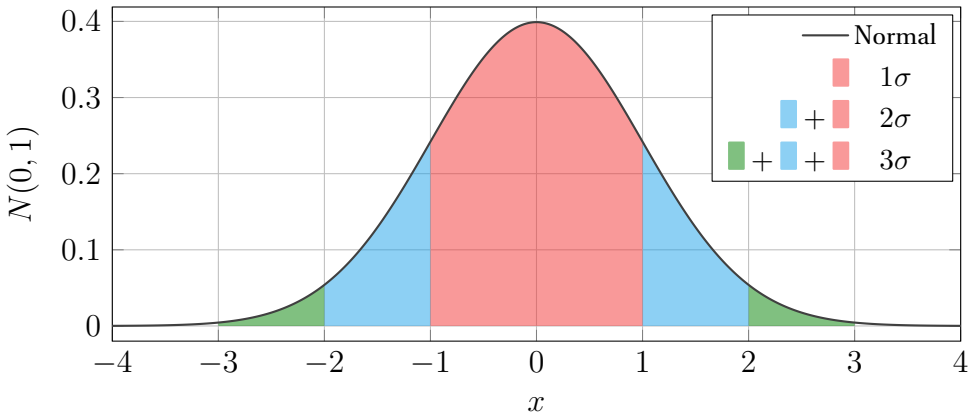


Figure 18: Plot of a standard normal distribution where the intervals associated with a 1σ, 2σ and 3σ deviation from the mean have been coloured in.

2.12 Bayes' theorem

An elementary concept in probability theory is the joint probability of multiple events occurring together. For example, a joint distribution $p(x, y)$ quantifies the probability that two events x and y both occur together. To retain the interpretation of p being a probability density, one of course has to require normalisation and non-negativity as dictated by the Kolmogorov axioms. A slightly more subtle point is that one also has to require

$$\int \mathrm{d}x\, p(x, y) = p(y) \qquad \text{and} \qquad \int \mathrm{d}y\, p(x, y) = p(x) \tag{2.172}$$

which is usually referred to as marginalisation. A set of events $x_1, \ldots, x_n$ are said to be independent whenever the joint probability p factorises into separate probability distributions $p_1, \ldots, p_n$, which describe the individual probabilities for $x_1, \ldots, x_n$ respectively, i.e.

$$p(x_1, \ldots, x_n) = p_1(x_1) \ldots p_n(x_n). \tag{2.173}$$

If, additionally, the events share the same probability distribution $p_1 = p_2 = \ldots = p_n$, they are said to be independent and identically distributed (iid). Of course, this definition generalises to any finite number of events.

Apart from the concept of joint probability, there is also the concept of a conditional probability. The probability of an event x, given that A is already known to be true is then denoted $p(x \mid A)$. Independent of the condition A one still require normalisation over the variable x:

$$\int \mathrm{d}x\, p(x \mid A) = 1 \qquad \forall A. \tag{2.174}$$

Crucially, the notions of conditional and joint probability are linked via Bayes' Theorem which states for all events x, y

$$\mathbb{P}(x, y) = \mathbb{P}(x \mid y)\, \mathbb{P}(y) = \mathbb{P}(y \mid x)\, \mathbb{P}(x) \tag{2.175}$$

or, equivalently,

$$\mathbb{P}(x \mid y) = \frac{\mathbb{P}(y \mid x)\, \mathbb{P}(x)}{\mathbb{P}(y)}. \tag{2.176}$$

For joint probabilities of more than two events, Bayes' theorem takes the form

$$\mathbb{P}(x, y, z) = \mathbb{P}(x \mid y, z)\, \mathbb{P}(y \mid z)\, \mathbb{P}(z) \tag{2.177}$$

which is also known as the chain rule of probabilities and generalises analogously for larger numbers of joint events.

In order to distinguish between these different probability distributions more clearly in an applied setting, many authors adopt the notation

$$p(\theta \mid \mathrm{data}) = \frac{L(\mathrm{data} \mid \theta)\, \pi(\theta)}{p(\mathrm{data})}. \tag{2.178}$$

A heavily used terminology in Bayesian inference is that of priors and posteriors. In the case of the previous equation, $\pi(\theta)$ is referred to as the prior distribution whereas the resulting conditional distribution $p(\theta \mid \mathrm{data})$ is called the posterior distribution. Moreover, $L(\mathrm{data} \mid \theta)$ is the likelihood and $p(\mathrm{data})$ is sometimes referred to as the Bayesian evidence.

In practice, it is often infeasible to calculate the Bayesian evidence $p(\mathrm{data})$ directly. However, it is possible to rewrite it as

$$p(\mathrm{data}) = \int \mathrm{d}\theta\; p(\mathrm{data}, \theta) = \int \mathrm{d}\theta\, L(\mathrm{data} \mid \theta)\, \pi(\theta) \tag{2.179}$$

using equations (2.172) and (2.175).

A subtle detail that is obscured by the above notation and occasionally glossed over in literature is the fact that particularly the likelihood $L(\mathrm{data} \mid \theta)$ already assumes that the given model is correct, i.e. that the observed data follows the assumed model for some parameter configuration.

Concretely, the chain rule of probabilities dictates

$$\mathbb{P}(\text{data}, \theta, \text{model}) = L(\text{data} \,|\, \theta, \text{model})\, \pi(\theta \,|\, \text{model})\, \underbrace{\mathbb{P}(\text{model})}_{=1 \text{ by assumption}} \tag{2.180}$$

which, given $\mathbb{P}(\text{model}) \neq 0$, results in

$$\mathbb{P}(\theta \,|\, \text{data}, \text{model}) = \frac{L(\text{data} \,|\, \theta, \text{model})\, \pi(\theta \,|\, \text{model})}{\mathbb{P}(\text{data} \,|\, \text{model})} \tag{2.181}$$

which more accurately reflects the underlying assumptions.[35]

In the frequentist view, it is assumed that all data points are samples from an existing true distribution underlying all observations. Subsequently, a probability density $p(x)$ is interpreted as the limit distribution of relative frequency of events x as the number of data points $N \longrightarrow \infty$. Thus, by repeating an experiment a large number of times, one should expect that the relative frequency of events resembles the underlying true distribution more and more accurately.

Conversely, in the Bayesian world view any observed data point exists in isolation and is not thought to originate from an underlying true distribution. A probability density $p(x)$ has the interpretation of quantifying the degree of belief that a new event x might occur in a future observation. While it is more laborious to conduct Bayesian regression than using the comparatively simpler frequentist estimators, its ability to systematically include information from earlier experiments or theoretical considerations via a prior distribution makes the Bayesian framework significantly more powerful.

For more details on the interpretation of Bayes' theorem and the difference between frequentist and Bayesian statistics, see e.g. [14, 68].

2.13 Important Probability Distributions

2.13.1 The Multivariate Normal Distribution

The normal distribution has a probability density function given by

$$p(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^N \det(\Sigma)}} \exp\left(-\frac{1}{2}\,(x-\mu)^a\, (\Sigma^{-1})_{ab}\, (x-\mu)^b\right) \tag{2.182}$$

where the vector $\mu \in \mathbb{R}^N$ represents the mean or location of the distribution while the matrix Σ determines the covariance between the components of x. In order for p to remain normalisable, the covariance matrix Σ must be positive definite. Using the differential Shannon entropy (see section 2.14), it is straightforward to show that the multivariate normal distribution maximises the entropy for fixed mean and covariance.

[35] While it is strictly only necessary that $\mathbb{P}(\text{model}) \neq 0$, one might argue from a philosophical point of view that only one true model can exist. In this context, model denotes an element of the abstract model function space and does not refer to any particular parametrisation of said model function.

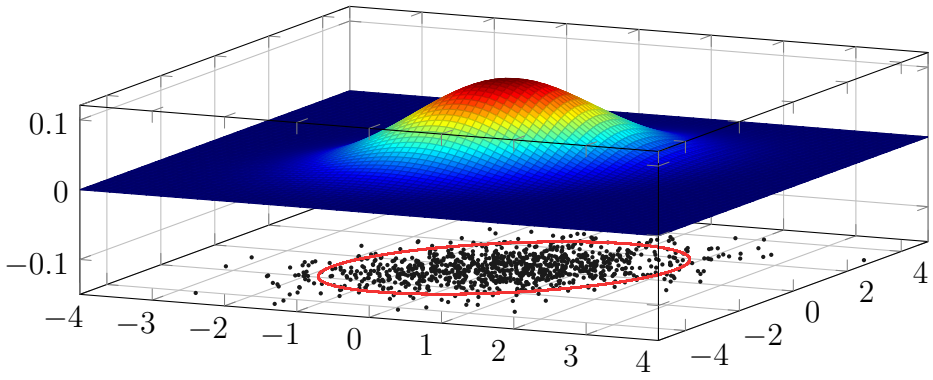


Figure 19: Illustration of random samples drawn from a multivariate normal distribution with covariance matrix $\Sigma = \left(\begin{smallmatrix} 1 & 1/2 \\ 1/2 & 2 \end{smallmatrix}\right)$ where the 2σ ellipse has been drawn in.

Due to the central limit theorem, the univariate and multivariate normal distributions are often considered the most important probability distributions for the experimental sciences. In essence, the central limit theorem asserts that for a set of N independent random vectors $X_1, ..., X_N$ which are individually distributed according to probability density functions $p_1, ..., p_N$ with common mean μ and common covariance matrix Σ the sample mean $m = \sum_{i=1}^{N} X_i/N$ is asymptotically distributed, up to order $O(1/\sqrt{N})$, as

$$\sqrt{N}\,(m - \mu) \xrightarrow{d} Y \sim N(0, \Sigma) \qquad (\text{as } N \longrightarrow \infty). \tag{2.183}$$

In words, this means that for large samples, where the individual contributions of each vector become insignificant and therefore fluctuations in the sample mean are smoothed, the (appropriately scaled) sample mean will always be normally distributed under the assumption that the probability distributions from which the samples were drawn have a common mean and common (finite) variance. The deviation of the distribution of the scaled sample mean $\sqrt{N}\,(m - \mu)$ from a perfect normal distribution decreases proportional to $O(1/\sqrt{N})$ for the number of samples N.

2.13.2 The Cauchy Distribution

The Cauchy distribution, which is a special case of the student's t-distribution for $\nu = 1$ degrees of freedom, is given by

$$p(x; \mu, s) = \text{Cauchy}(x; \mu, s) \coloneqq \frac{1}{\pi}\,\frac{s}{s^2 + (x - \mu)^2}. \tag{2.184}$$

The Cauchy distribution is subject to a location parameter μ and a scale parameter s, which controls the width of the distribution, similar to the standard deviation for the normal distribution. It is straightforward to calculate that its cumulative distribution function must be given by

$$F(x; \mu, s) = \int_{-\infty}^{x} \mathrm{d}x'\, p(x'; \mu, s) = \frac{1}{\pi} \int_{-\infty}^{(x-\mu)/s} \mathrm{d}u\, \frac{1}{1 + u^2} = \frac{1}{\pi} \arctan\left(\frac{x - \mu}{s}\right) + \frac{1}{2}. \tag{2.185}$$

Compared with a normal distribution, the density of the Cauchy distribution is not peaked as sharply. Also, its tails decay to zero at a much slower rate than the tails of a normal distribution as shown in figures 20 and 21.

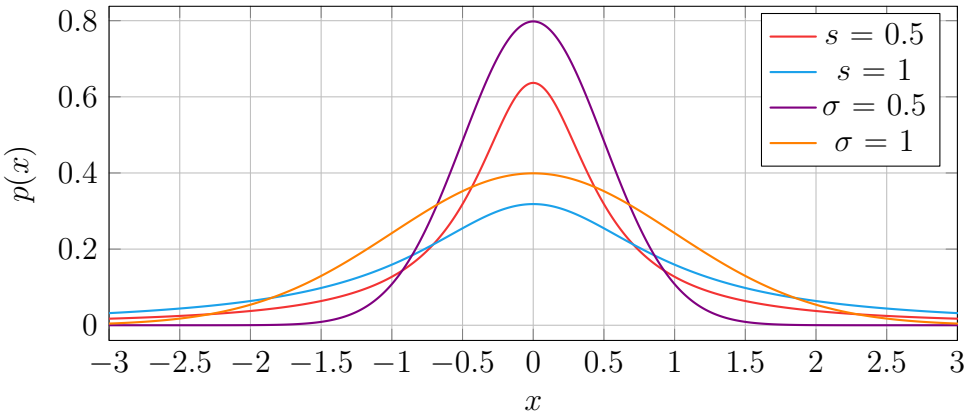


Figure 20: Plot of Cauchy and normal distributions at location $\mu = 0$ for different values of the scale parameter s and the standard deviation σ, respectively. One can see that in the case $\sigma = s$, the probability volume of a normal distribution is much more localised than a Cauchy distribution, which has much fatter tails in comparison.

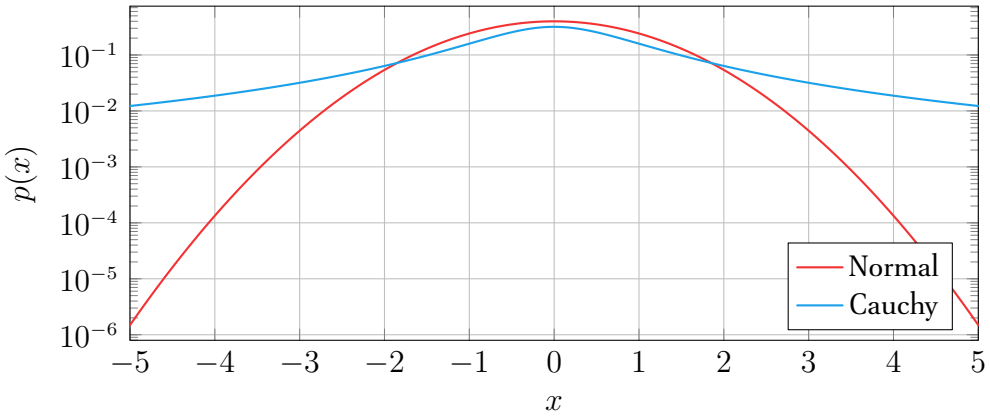


Figure 21: This logarithmic plot emphasises the difference in drop-off rate between a standard normal distribution and a standard Cauchy distribution, whose tails decay much less rapidly and therefore contain a significantly larger portion of the total probability volume.

2.13.3 The χ^2-Distribution

The χ^2 distribution is a probability distribution which depends on a single parameter $k \in \mathbb{N}$ that is usually referred to as the degrees of freedom. It frequently arises in the description of natural phenomena due to its relation to the normal distribution:

Given a set of k random variables $Z_1, ..., Z_k$ that are mutually independent and identically distributed (iid) according to the standard normal distribution $N(0, 1)$, the sum of their squares

will be distributed according to χ^2_k. That is,

$$Z_1, ..., Z_k \sim N(0,1) \qquad \Longrightarrow \qquad \left(\sum_{j=1}^{k} Z_j^2\right) \sim \chi^2_k. \tag{2.186}$$

The probability density function of the χ^2_k distribution with $k \in \mathbb{N}$ can be shown to be

$$\chi^2_k(x) := \frac{x^{\frac{k}{2}-1}\, e^{-\frac{x}{2}}}{2^{k/2}\,\Gamma(k/2)}\,\Theta(x) \tag{2.187}$$

with $\Theta(x)$ the Heaviside function to ensure that $\chi^2_k(x) = 0$ for all $x < 0$.[36] The χ^2 distribution can be found as the probability density which maximises the Shannon entropy functional under the constraints

$$\mathbb{E}(x) = k, \quad \mathbb{E}\Big(\ln(x)\Big) = \psi\left(\frac{k}{2}\right) + \ln(2) := \frac{\Gamma'(k/2)}{\Gamma(k/2)} + \ln(2) \tag{2.188}$$

where ψ represents the digamma function. It can further be shown that

$$\frac{\chi^2_k - k}{\sqrt{2k}} \sim N(0,1) \qquad (\text{as } k \longrightarrow \infty) \tag{2.189}$$

from which one may read off that the mean of a χ^2_k distribution is given by k while its variance is given by $2k$.

2.13.4 The Generalised Student's t-Distribution

The generalised t-distribution has a continuous density function given by

$$p(y;\mu,\sigma,\nu) = \frac{\Gamma(\frac{1+\nu}{2})}{\Gamma(\frac{\nu}{2})}\,\frac{1}{\sqrt{\pi\nu\sigma^2}}\left(1 + \frac{1}{\nu}\left(\frac{y-\mu}{\sigma}\right)^2\right)^{-\frac{1+\nu}{2}} \tag{2.190}$$

where the $\mu \in \mathbb{R}$ is the mean, $\frac{\nu}{\nu-2}\,\sigma^2$ is the variance and $\nu \in \mathbb{N}$ is the so-called degrees of freedom parameter.[37] The resemblance this expression bears to the density function of the normal distribution is not just coincidental—in the limit of infinite degrees of freedom $\nu \longrightarrow \infty$ the student's t-distribution recovers the normal distribution exactly. For $\nu = 1$, one obtains exactly the Cauchy distribution. Therefore, any standard t-distribution is enveloped by the standard Cauchy and normal distributions in figure 20.

Given a set of k independent random variables, $Z_1, ..., Z_k$ that are drawn as samples from a normal distribution $N(\mu,\sigma)$, the true mean μ and true variance σ^2 can be estimated using the

[36] Note that for $k = 1$, $\chi^2_1(x)$ diverges as $x \longrightarrow 0$.
[37] The mean of this distribution is only defined for $\nu > 1$ and the variance exists only when $\nu > 2$.

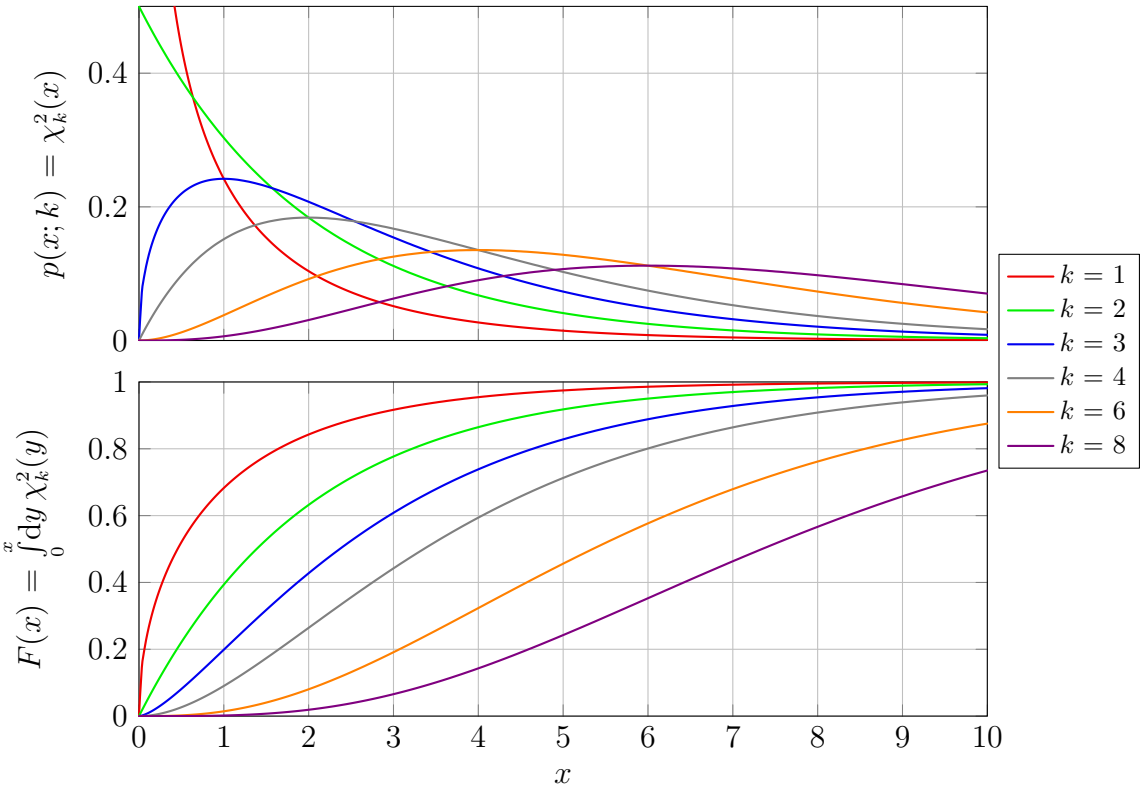


Figure 22: Plots of the probability density function (on top) and the cumulative distribution function (on the bottom) of the χ^2_k distribution for different degrees of freedom k.

sample mean $\bar{Z}$ and sample variance S^2 which are respectively defined by

$$\bar{Z} = \frac{1}{k}\sum_{i=1}^{k} Z_i \qquad \text{and} \qquad S^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\bar{Z} - Z_i)^2. \tag{2.191}$$

While the quantity $(\bar{Z} - \mu)/(\sigma/\sqrt{k})$ is of course distributed according to a standard normal distribution $N(0,1)$, the mean and variance are generally not known. However, it can be shown that the quantity $(\bar{Z} - \mu)/(S/\sqrt{k})$ is distributed according to a standard t-distribution with $\nu = k - 1$ degrees of freedom. Therefore, the t-distribution is particularly useful for the analysis of small datasets. Multivariate extensions of the student's t-distribution are discussed in [27, 28, 50].

2.13.5 The Snedecor F-Distribution

The probability density of the Snedecor F-distribution is given by

$$p(x; d_1, d_2) = \frac{\Gamma\left(\frac{d_1}{2} + \frac{d_2}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} \left(1 + \frac{d_1}{d_2}\, x\right)^{-\frac{d_1+d_2}{2}} x^{\frac{d_1}{2}-1}\, \Theta(x) \tag{2.192}$$

where $d_1, d_2 \in \mathbb{N} \setminus \{0\}$ are two parameters describing the degrees of freedom. Analogous to how the sum of squares of random variables following a standard normal distribution can be shown to be distributed according to χ^2, it can be shown that a random variable distributed according to a

standard t-distribution with ν degrees of freedom fulfils

$$X \sim t(\nu) \qquad \Longrightarrow \qquad X^2 \sim F(1, \nu). \tag{2.193}$$

In addition, the F-distribution also satisfies

$$X \sim F(d_1, d_2) \qquad \Longrightarrow \qquad X^{-1} \sim F(d_2, d_1). \tag{2.194}$$

Finally, since the t-distribution recovers the normal distribution in the limit as $\nu \longrightarrow \infty$, it can also be shown that the F-distribution recovers the χ^2-distribution via

$$X \sim F(k, N-k) \qquad \Longrightarrow \qquad \lim_{N\to\infty} kX = Y \sim \chi^2_k. \tag{2.195}$$

2.14 Information Entropies

The modern information-theoretic formulation of statistics postulates that the amount of prior knowledge built into a probability distribution should be minimal. That is to say, the information content per measurement should be maximal. In this context, information is a precisely defined quantity which is generally measured by some choice of an entropy functional.

The (discrete) Shannon entropy H is defined as

$$\mathrm{H}(X) = -\sum_{x\in X} p(x) \ln\Big(p(x)\Big) \tag{2.196}$$

where $X = \{x_1, ..., x_N\}$ is some set of events. It can be shown that the discrete Shannon entropy is unique under a set of technical assumptions on the properties of the entropy including additivity and an inverse proportionality to the probability of events x_i. In order to aid the interpretation of the Shannon entropy as a measure of information, practitioners of information theory often choose the logarithm of base 2 in the Shannon entropy such that the resulting value is in units of bits.

Naïvely, one might generalise this to the so-called differential Shannon entropy S via

$$S[p] \coloneqq \mathbb{E}\Big(-\ln(p)\Big) = -\int \mathrm{d}x\, p(x)\, \ln\Big(p(x)\Big). \tag{2.197}$$

However, there are two immediate problems with this expression: the differential Shannon entropy $S[p]$ does not actually correspond to the limit of the discrete Shannon entropy H for $|X| \longrightarrow \infty$. Instead, it can be shown that it differs from this limit of the discrete Shannon entropy by an infinite offset. Moreover, the differential Shannon entropy is not invariant under reparametrisation of the

random variable. That is,

$$S[p] = \mathbb{E}_x\Big(-\ln\big(p(x)\big)\Big) = -\int \mathrm{d}x\, p(x)\, \ln\big(p(x)\big) \tag{2.198}$$

$$= -\int \mathrm{d}y \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| p(y) \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right| \ln\left(p(y)\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|\right) \tag{2.199}$$

$$= \underbrace{-\int \mathrm{d}y\, p(y)\, \ln\big(p(y)\big)}_{\mathbb{E}_y\left(-\ln\left(p(y)\right)\right)} \underbrace{-\int \mathrm{d}y\, p(y)\, \ln\left(\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|\right)}_{\neq 0} \neq \mathbb{E}_y\Big(-\ln\big(p(y)\big)\Big). \tag{2.200}$$

While it is the most widely used choice of entropy functional, there are also various extensions to the Shannon entropy which have been proposed, such as the Rényi entropy S_α given by

$$S_\alpha[p] := \frac{1}{\alpha - 1} \ln\left(\int \mathrm{d}x\, p(x)\, p^{\alpha-1}(x)\right) \tag{2.201}$$

where α is called the order parameter of the Rényi entropy (see [60]). In the limit $\alpha \longrightarrow 1$, the Rényi entropy reverts back to the Shannon entropy under application of L'Hôpital's rule.

Given that the definition of absolute entropy is problematic in the differential setting, one instead considers the relative entropy between probability distributions of random variables. The relative Shannon entropy can be shown to correspond precisely to the Kullback–Leibler divergence discussed in section 2.15.

The modern view of information theory is typically that the Kullback–Leibler divergence $D_{\mathrm{KL}}[p : q]$ is the fundamental object from which the differential Shannon entropy $S[p]$ is obtained by comparing a given probability distribution p to a uniform probability distribution q. Finally, the discrete Shannon entropy is recovered as the discretised form of the Kullback–Leibler divergence, again by comparing a distribution p against a uniform distribution $q = 1/N$.

2.15 Obtaining the Fisher Metric from Information Divergences

An information divergence function is a measure of dissimilarity between two probability distributions. However, information divergences provide less structure than a metric function due to the fact that the requirements posed on them are much weaker. Specifically, they are generally not symmetric and fail to satisfy a triangular inequality.

However, this comparative lack of structure also facilitates their increased versatility compared with metric functions: They are not restricted to use within any given family of probability distributions but instead allow for a systematic comparison of arbitrary probability distributions which share a common domain. Moreover, the this section outlines how information divergences can be used to induce metric tensor fields.

In information theory, a function $D[p : q]$ is called an information divergence[38] if it satisfies for

[38] The purpose of the colon notation $D[p : q]$ used by Amari and others is only to emphasise the asymmetry of the divergence function, i.e. that the order of arguments is usually not interchangeable.

all probability distributions p and q with common domain:

$$\text{(Positivity)} \qquad 0 \leq D[p : q] \tag{2.202}$$

$$\text{(Definiteness)} \qquad 0 = D[p : q] \iff p = q. \tag{2.203}$$

Clearly, these comparatively weak requirements result in a very large set of admissible divergence functions. In addition to the above requirements of positivity and definiteness, the definition of an information divergence used by Amari in [2] adds the requirement that $D[p : q]$ should be analytic, i.e. that an information divergence should be locally representable by a convergent power series. Furthermore, attention is usually restricted to families of probability distributions which can be characterised using a finite number of parameters which vary continuously over some domain in $\mathbb{R}^n$, which is the case in almost all practical applications.

A prominent class of divergence functions is given by the so-called f-divergences D_f defined by

$$D_f[p : q] \coloneqq \int \mathrm{d}y\, p(y)\, f\left(\frac{q(y)}{p(y)}\right) \tag{2.204}$$

where f is a differentiable convex function satisfying $f(1) = 0$ and is consequently said to generate D_f. It is straightforward to check that such a D_f indeed satisfies the criteria for divergence functions. A special case of an f-divergence is the Kullback–Leibler divergence, denoted D_{KL}, where the generating function is chosen as $f(u) = -\ln u$ such that

$$D_{\mathrm{KL}}[p : q] \coloneqq \int \mathrm{d}y\, p(y)\, \ln\left(\frac{p(y)}{q(y)}\right). \tag{2.205}$$

Using the Kullback–Leibler divergence, one obtains the Fisher metric g as its Hessian via

$$g_{ab}(\theta) = \left[\frac{\partial^2}{\partial\psi^a\, \partial\psi^b}\, D_{\mathrm{KL}}\big[p(y;\theta) : p(y;\psi)\big]\right]_{\psi=\theta}. \tag{2.206}$$

Given the convexity of the generating function f, it is not hard to see that the Hessian of the Kullback–Leibler divergence is positive definite as long as the involved probability densities are suitably differentiable. Moreover, it can be shown that the same metric is obtained from any f-divergence.[39] Therefore, instead of the Kullback–Leibler divergence, one is free to choose any other f-divergence in equation (2.206). Provided that the integrand of the Kullback–Leibler divergence is regular enough such that one may exchange the order of integration and differentiation, the Fisher metric can also be obtained by calculating the Hessian of the integrand first and evaluating the integral afterwards.

In general, it is quite difficult to find an exact analytical expression for the the Kullback–Leibler divergence between two given probability distributions, especially so in the multivariate case.

[39] For example, equally valid choices of divergence-generating functions include $f = t^2 - 1$ and $f = |t - 1|$.

However, for two multivariate normal distributions it is straightforward to work out that

$$D_{\mathrm{KL}}\Big[p(x;\mu_0,\Sigma_0) : p(x;\mu,\Sigma)\Big] = \frac{1}{2}\left[\ln\left(\frac{\det(\Sigma)}{\det(\Sigma_0)}\right) - n + \operatorname{tr}\Big(\Sigma^{-1}\,\Sigma_0\Big) + \mu_\delta^\top\,\Sigma^{-1}\,\mu_\delta\right] \tag{2.207}$$

where $\mu = \mu_0 + \mu_\delta$ and $\Sigma = \Sigma_0 + \Sigma_\delta$. As a special case, the divergence between univariate normal distributions is then given by

$$D_{\mathrm{KL}}[p_1 : p_2] = \int_{-\infty}^{\infty} \mathrm{d}x\, p_1(x)\, \ln\left(\frac{p_1(x)}{p_2(x)}\right) = \ln\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2}\left(\left(\frac{\sigma_1}{\sigma_2}\right)^2 + \left(\frac{\mu_1-\mu_2}{\sigma_2}\right)^2 - 1\right). \tag{2.208}$$

A useful property for the calculation of f-divergences between univariate location-scale distributions derived in [53] is

$$D_f[\mu_1, s_1 : \mu_2, s_2] = D_f\left[0, 1 : \frac{\mu_2-\mu_1}{s_1}, \frac{s_2}{s_1}\right] = D_f\left[\frac{\mu_1-\mu_2}{s_2}, \frac{s_1}{s_2} : 0, 1\right]. \tag{2.209}$$

As mentioned in section 2.14, the Kullback–Leibler divergence quantifies the relative entropy between probability distributions and is associated specifically with the Shannon entropy. Specifically, $D_{\mathrm{KL}}[p : q]$ can be interpreted as the amount of information which is lost (i.e. the increase in entropy) by approximating the probability distribution p through the distribution q. However, in contrast to the differential Shannon entropy the Kullback–Leibler divergence is invariant with respect to reparametrisation of the random variable which one can verify as follows: a univariate probability density function can be reparametrised according to

$$p(x)\,\mathrm{d}x = p(y)\,\mathrm{d}y \qquad \Longrightarrow \qquad p(y) = p(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \tag{2.210}$$

which generalises in the multivariate case to

$$p(x)\,\mathrm{d}x = p(y)\,\mathrm{d}y \qquad \Longrightarrow \qquad p(y) = p(x)\left|\det\left(\frac{\mathrm{d}x}{\mathrm{d}y}\right)\right| = p(x)\left|\det\left(\frac{\partial x^a}{\partial y^b}\right)\right|. \tag{2.211}$$

Inserting this into the Kullback–Leibler divergence while transforming the measure accordingly yields

$$\int \mathrm{d}y\, p(y)\, \ln\left(\frac{p(y)}{q(y)}\right) = \int \mathrm{d}x \left|\det\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)\right| p(x) \left|\det\left(\frac{\mathrm{d}x}{\mathrm{d}y}\right)\right| \ln\left(\frac{p(x)\left|\det\left(\frac{\mathrm{d}x}{\mathrm{d}y}\right)\right|}{q(x)\left|\det\left(\frac{\mathrm{d}x}{\mathrm{d}y}\right)\right|}\right) \tag{2.212}$$

$$= \int \mathrm{d}x\, p(x)\, \ln\left(\frac{p(x)}{q(x)}\right). \tag{2.213}$$

An illustration of the Kullback–Leibler divergence between two normal distributions is given in figure 23.

Coming from the point of view of information divergences, it seems rather arbitrary to define the

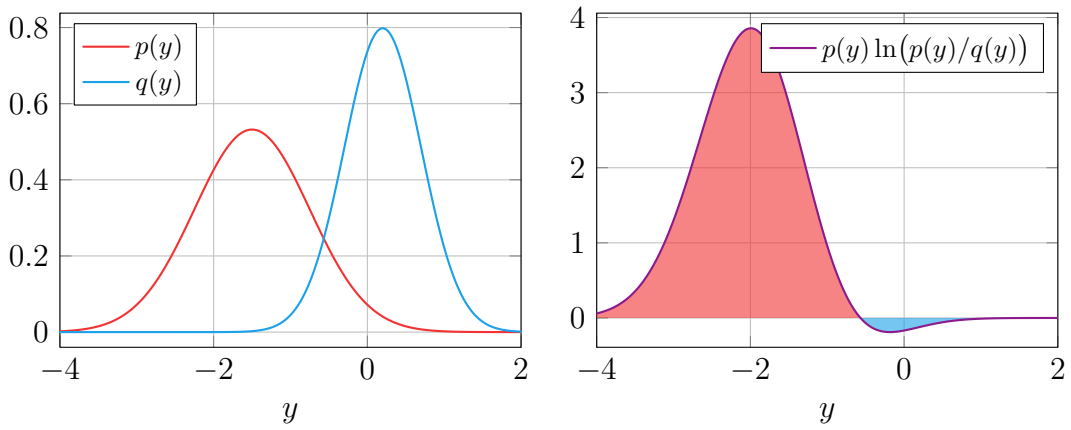


Figure 23: The left-hand side depicts two normal distributions p and q. The right-hand side shows a plot of the function $p\ln\big(p/q\big)$ which means that the Kullback–Leibler divergence $D[p:q]$ is then given by the area under this curve.

Fisher metric as an approximation to the Kullback–Leibler divergence (or f-divergences in general), since one might equally well choose some other information divergence, resulting in a different metric. For this reason, the definition of the Fisher metric via the Kullback–Leibler divergence (or any other f-divergence) should not be considered as the fundamental definition of the Fisher metric.

Instead, it was shown by Čencov that there exists a unique choice of metric which remains invariant under a family of probabilistically meaningful mappings which are referred to as "congruent embeddings via Markov morphisms" (see [79]). Intuitively, these Markov morphisms may be thought of as facilitating a coarse-graining of the available data. While the original proof by Čencov is formulated in the language of category theory, a simplified and more geometric proof of this theorem is given in [16, 41]. From the geometric proof, it is in particular observed that for normalised models (i.e. probability distributions), the unique metric identified by Čencov coincides precisely with the metric that is obtained from the Hessian of the Kullback–Leibler divergence, i.e. through equation (2.206).

In conclusion, one should consider the formula for the calculation of the Fisher metric from the Kullback–Leibler divergence as little more than a convenient recipe whereas the fundamental reason for this recipe lies in the fact that this is the unique metric which exhibits the aforementioned desirable behaviour under coarse-graining of the observed data, as guaranteed by Čencov's theorem.

3 The Geometric Framework of Parameter Inference

With the introduction of the core concepts of differential geometry and probability theory out of the way, this section focuses on how the toolkit and language of differential geometry can be used to study statistical problems.

3.1 Datasets and Error Distributions

In general, a dataset consists of three parts: observations $y_i \in \mathcal{Y}$, the conditions $x_i \in \mathcal{X}$ at which the respective observations y_i were made, and a specification of the error distributions associated with the observations y_i. Typically, the spaces $\mathcal{X}$ and $\mathcal{Y}$ where the observations and the associated conditions take their values in is (some convex cone of) $\mathbb{R}^d$. For example, when measuring physical quantities such as absolute temperature, negative values are disallowed for physical reasons, i.e. $\mathcal{Y} = \mathbb{R}^+_0$.

The error distribution must be supplied "by hand" through the experimenter and it encapsulates properties of the measurement such as systematic errors, statistical noise and possibly bias.[40] Apart from the fact that the error distribution should be a probability distribution that is defined on all of $\mathcal{Y}$ and is sufficiently differentiable, no further conceptual requirements are necessary in order to be able to use the differential geometric formalism.

Since it is very convenient to work with, the error distribution is almost always chosen to be the normal distribution. As also discussed in section 2.13.1, the central limit theorem is generally invoked as a theoretical justification for this choice. However, for small datasets where the individual errors associated with the observations are hard to quantify, it may instead be advisable to pick an error distribution which contains a larger proportion of the total probability volume in its tails and thus tends to overestimate the error spread. For example, it may be more appropriate to model the errors using a Cauchy distribution (or any other student's t-distribution) in place of a normal distribution in such cases.

Using a multivariate error distribution with non-diagonal covariance matrix, one can also encode the fact that the individual observations in the dataset were not made independently of each other. That is, a diagonal covariance matrix corresponds to independent measurements.

Although this is not discussed in the scope of this thesis, it would be possible to further adapt this formalism to account for errors in the conditions x_i as well. One way to deal with such errors has been proposed for example in [30, 31]. Essentially, it consists of adding a suitable correction term to the Fisher metric which incorporates a linear approximation of the sensitivity of the model to the observations $x \in \mathcal{X}$.

The main distinction between what is considered an observation and and what is considered an observation condition is typically made by the model (see section 3.2). The observation conditions

[40] Although experimenters generally like to make a distinction between measurement errors and uncertainty, these terms will be used synonymously and interchangeably throughout this thesis.

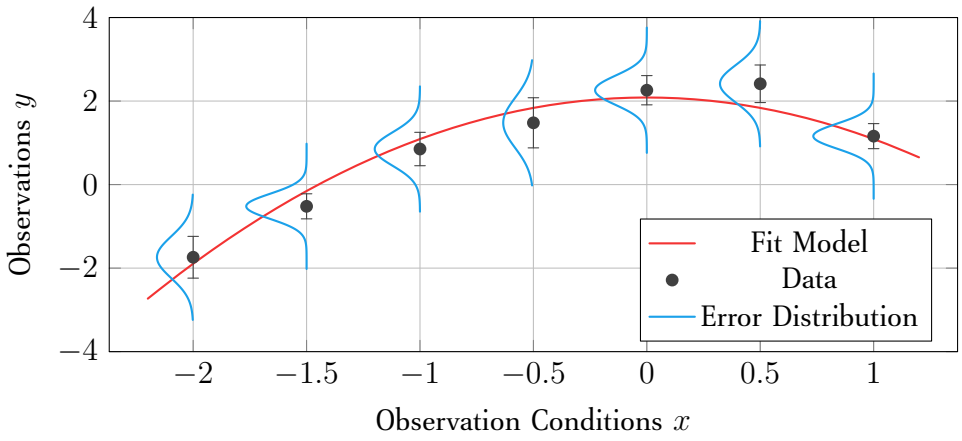


Figure 24: Visualisation of normal error distributions for independent observations in the case $\dim \mathcal{X} = 1 = \dim \mathcal{Y}$. Of course, the error distributions should be thought of as perpendicular to the x-y plane.

correspond to an input that is made to the model map whereas the observations are quantities that can be predicting using the given model.

A notation which is heavily used throughout this thesis is

$$y_{\text{data}} \coloneqq (y_1, \dots, y_N) \in \mathcal{D} = \mathcal{Y}^N \tag{3.1}$$

i.e. one considers the ordered collection of all observations in a dataset to form a single data point $y_{\text{data}} \in \mathcal{D} = \mathcal{Y}^N$ rather than considering the observations $y_i \in \mathcal{Y}$ individually. This does not exclude the possibility that each individual observation has multiple components, i.e. it is possible that $\dim \mathcal{Y} > 1$ in which case y_{data} has $N \dim \mathcal{Y}$ components. Similarly, one might abbreviate the conditions as

$$x_{\text{data}} \coloneqq (x_1, \dots, x_N) \in \mathcal{X}^N. \tag{3.2}$$

Although the above definition of a dataset as containing observations $y \in \mathcal{Y}$ and observation conditions $x \in \mathcal{X}$ can be applied to most situations, some experiments may involve data collection where observation conditions cannot be collected. As an example, one might make repeated observations without being able to influence any of the conditions under which these measurements are taken. Such data can be cast into an $\mathcal{X}$-$\mathcal{Y}$ form for example by creating a normalised histogram of the observations with appropriate bin sizes and considering the horizontal location of the bins as the conditions and their heights as the observations. Consequently, the histogram can be described by also choosing a probability distribution as the model itself.

3.2 Models

Generally, a model is a map $y_{\text{model}} : \mathcal{X} \times \mathcal{M} \longrightarrow \mathcal{Y}$, which takes an observation condition $x \in \mathcal{X}$ and a parameter configuration $\theta \in \mathcal{M}$ and maps them to a predicted observation $y \in \mathcal{Y}$. Further, it is assumed that y_{model} is sufficiently differentiable with respect to the parameters $\theta \in \mathcal{M}$. Since the Riemann tensor contains second derivatives of the metric, which in turn can be written in

terms of the first derivatives of the model map y_{model} with respect to its parameters, this means that $y_{\text{model}}(x;\,\cdot\,) : \mathcal{M} \longrightarrow \mathcal{Y}$ should ideally be at least three times continuously differentiable if one aims to discuss curvature (cf. equations (2.105) and (3.17)).

On the other hand, no such requirement is necessary for the model map with respect to the observation conditions $x \in \mathcal{X}$ from the perspective of the mathematical formalism. Since real datasets only ever contain a finite number of data points, the model map $y_{\text{model}}(x;\theta)$ is only sampled at finitely many conditions $x \in \mathcal{X}$. For this reason, it is conceivable that the model map could even be discontinuous with respect to $x \in \mathcal{X}$ in principle. However, from a phenomenological perspective, it would be desirable for the model map $y_{\text{model}}(x;\theta)$ to be stable in the sense that a small perturbation in the observation conditions does not result in large (chaotic) changes in the predictions of the model. For this reason, while noting that this requirement could be strengthened or weakened, only models which are continuous with respect to the conditions $x \in \mathcal{X}$ are considered here, i.e. $y_{\text{model}}(\,\cdot\,;\theta) \in C^0(\mathcal{X},\mathcal{Y})$.

Given this notion of a model map, there are a large numbers of ways in which such a model could be specified in practice: the simplest option is of course to provide an explicit analytic expression. On the other hand, one might prescribe the model piecewise either with respect to the conditions $x \in \mathcal{X}$ or the parameters $\theta \in \mathcal{M}$ as so long as the differentiability with respect to the parameters is still given. Alternatively, a model can be specified implicitly, for example in the form of a system of differential equations

$$(\mathscr{D}_x\, y_{\text{model}})(x;\theta) = f\left(x, y_{\text{model}}, \frac{\partial}{\partial x^{a_1}}\, y_{\text{model}}, \ldots, \frac{\partial}{\partial x^{a_1}} \cdots \frac{\partial}{\partial x^{a_m}}\, y_{\text{model}}; \theta\right) \tag{3.3}$$

where $\mathscr{D}_x$ is some differential operator depending on up to m-th derivatives of the model with respect to the conditions x. This is a use case which is frequently encountered in phenomenological disciplines such as systems biology where it is often too cumbersome—or even impossible—to express the model in closed analytical form.

In principle, even an artificial neural network which can be comprised of millions of neurons and have billions of adjustable parameters fits this definition. That is, it takes in a set of inputs $x \in \mathcal{X}$ and non-linearly maps them to a set of outputs using its pre-trained weights. Incidentally, the universal approximation theorem[41] from which neural network methods derive their great versatility and power is also responsible for large amounts of data that are necessary to determine the appropriate weights.

The high malleability of the model map associated with a neural network demonstrates that no structure whatsoever is hard-wired into the neural network in the sense that its shape is purely dictated by the values of its parameters. In comparison, a polynomial model function $y_{\text{model}}(x;a,b,c) = ax^2 + bx + c$ has a very rigid structure in the sense that there is no parameter configuration (a,b,c) with which it could faithfully reproduce the prediction of a polynomial whose degree is larger than three.

[41] The universal approximation theorem guarantees that a neural network can be used to approximate any continuous function arbitrarily well if given a sufficient number of adjustable weight parameters.

In order for a model to be predictive, its number of parameters—while it may be very large—must remain finite to be able to infer their values from measurement. In physical models, the parameters usually correspond to fundamental constants, such as Newton's gravitational constant or the charge of the electron.

Equivalent notations which are used interchangeably to denote the model map are

$$y_{\text{model}}(x;\theta) \equiv y(x;\theta) \equiv y_\theta(x). \tag{3.4}$$

3.3 The Likelihood Function

A central quantity in parameter inference is the likelihood function $L(\text{data}\,|\,\theta)$, which quantifies the probability of measuring a given dataset under the assumption that the dataset would be generated by a model y_{model} which depends on the parameters θ. Since any measurement in the real world is bound to feature noise and uncertainties due to the fact that one can only attain finite precision, there is of course an error distribution associated with each measurement, which the likelihood also takes into account.

Its most important use case is to determine the parameter configuration θ_{MLE} which maximises the likelihood function and (in precisely this sense) provides the best possible fit of the model y_{model} to the dataset. This process is known as maximum likelihood estimation (MLE). Also, one can test hypotheses by comparing their corresponding likelihood functions. Moreover, the Neyman–Pearson Lemma (see [51]) states that the likelihood ratio is the most powerful test when it comes to discriminating between so-called simple hypotheses.

In cases where the observations in a dataset of size N are mutually independent, the likelihood can be factored into a product of probabilities

$$L(\text{data}\,|\,\theta) \equiv L\Big(y_{\text{data}}\,\Big|\,y_\theta(x)\Big) = \prod_{j=1}^{N} \mathbb{P}_j\Big(y_j\,\Big|\,y_{\text{model}}(x_j;\theta)\Big) \tag{3.5}$$

where $\mathbb{P}_j$ denotes the error distribution associated with the j-th observation. With rare exceptions, it is more convenient to work with the log-likelihood $\ell := \ln \circ L$ in practice, instead of the bare likelihood L itself. For one, this effects that in the case of independent observations the product of probabilities in the likelihood is turned into a sum. Also, the rescaling provided by the logarithm drastically reduces the rounding errors of floating-point arithmetic during concrete numerical computations of the likelihood, since its value can often be very small.

Using the abbreviation $\zeta^a(\theta) := \Big(y_{\text{data}} - h(\theta)\Big)^a$ with $h : \mathcal{M} \longrightarrow \mathcal{D}$ the embedding map from equation (3.26), the likelihood for single-component observations (i.e. $\dim \mathcal{Y} = 1$) and a multivariate normal error distribution is given by

$$L(\text{data}\,|\,\theta) = \frac{1}{\sqrt{(2\pi)^N \det(\Sigma)}} \exp\left(-\frac{1}{2}\,\zeta^i(\theta)\,(\Sigma^{-1})_{ij}\,\zeta^j(\theta)\right) \tag{3.6}$$

such that the corresponding log-likelihood is

$$\ell(\text{data}\,|\,\theta) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\Big(\det(\Sigma)\Big) - \frac{1}{2}\,\zeta^i(\theta)\,(\Sigma^{-1})_{ij}\,\zeta^j(\theta) \tag{3.7}$$

where Σ denotes the covariance matrix of the dataset. Likelihoods of the this form are usually simply referred to as "normal likelihoods". If the observations are mutually independent, the covariance matrix is diagonal $\Sigma = \operatorname{diag}(\sigma_1, ..., \sigma_N)$ and the expression for the normal likelihood simplifies to

$$L(\text{data}\,|\,\theta) = \frac{1}{\sqrt{(2\pi)^N \prod_{j=1}^N \sigma_j}} \exp\left(-\frac{1}{2}\sum_{i=1}^N \left(\frac{y_i - y_\theta(x_i)}{\sigma_i}\right)^2\right). \tag{3.8}$$

The gradient of the log-likelihood $\mathrm{d}\ell$ with respect to the model parameters in some chart (U, θ) is a covector field given by

$$\mathrm{d}\ell = \frac{\partial \ell}{\partial \theta^j}\,\mathrm{d}\theta^j = \frac{1}{L}\frac{\partial L}{\partial \theta^j}\,\mathrm{d}\theta^j \tag{3.9}$$

and is often referred to as the score in literature. As calculated in equation (3.15), the expectation value of the components of the score vanishes.

Taking the ratio between two likelihoods allows for the direct comparison of so-called simple hypotheses

$$\Lambda(\text{data}\,|\,\theta_0 : \theta_1) \coloneqq \frac{L(\text{data}\,|\,\theta_0)}{L(\text{data}\,|\,\theta_1)} \tag{3.10}$$

where the hypothesis θ_0 is said to be more likely if $\Lambda < 1$ and the hypothesis θ_1 is said to be more likely if $\Lambda > 1$. If the distribution of the likelihood ratio as a random variable is known, this information can be used to determine whether one of the hypotheses can actually be rejected in favour of the other on the grounds of the available evidence.

Clearly, the difference of the log-likelihoods is just as effective at discriminating between the hypotheses as the likelihood ratio, i.e.

$$\lambda(\theta_0 : \theta_1) \coloneqq \ln\Big(\Lambda(\text{data}\,|\,\theta_0 : \theta_1)\Big) \coloneqq \ell(\text{data}\,|\,\theta_0) - \ell(\text{data}\,|\,\theta_1) \tag{3.11}$$

except that the threshold where one hypothesis becomes more likely than the other is now zero instead of one and the sign is reversed due to the logarithm.

To obtain an estimator for the parameter configuration θ_{MLE} which maximises the likelihood[42] one has the condition

$$(\mathrm{d}\ell)(\theta_{\text{MLE}}) \overset{!}{=} 0 \qquad \Longleftrightarrow \qquad \left.\frac{\partial \ell}{\partial \theta^a}\right|_{\theta_{\text{MLE}}} \overset{!}{=} 0. \tag{3.12}$$

Also, to ensure that θ_{MLE} truly constitutes a maximum of ℓ one must verify the negative-definiteness of its Hessian $\operatorname{Hess}_\ell(\theta_{\text{MLE}})$. While it is not without its limitations, the maximum likelihood method enjoys popularity in the scientific community mainly due to its versatility and due to the fact that

[42] Since the natural logarithm is strictly monotonically increasing, maximising the log-likelihood ℓ is equivalent to maximising the likelihood L.

the estimators derived from the maximum likelihood principle have desirable properties. That is, the estimators obtained from equation (3.12) are usually asymptotically efficient and consistent (see e.g. [45]).

Computationally, the most expensive step in the calculation of the likelihood is typically the evaluation of the model since its prediction must be calculated for every value data point.

3.4 The Fisher Metric

In section 2.15, it was established that the Fisher metric on spaces of probability distributions is obtained as the Hessian of the Kullback–Leibler divergence. Looking more closely at this recipe, one recognises that the Kullback–Leibler divergence can also be expressed as an expectation value. In particular, the divergence between two likelihood functions $L(y_{\text{data}} \,|\, \theta)$ and $L(y_{\text{data}} \,|\, \psi)$ results in

$$g_{ab}(\theta) = \left[\frac{\partial^2}{\partial\psi^a\,\partial\psi^b}\, D_{\text{KL}}\Big[L(y_{\text{data}} \,|\, \theta) \,:\, L(y_{\text{data}} \,|\, \psi)\Big]\right]_{\psi=\theta} = \left[\frac{\partial^2}{\partial\psi^a\,\partial\psi^b}\, \mathbb{E}_{y_{\text{data}}}\Big(\ell(\theta) - \ell(\psi)\Big)\right]_{\psi=\theta} \tag{3.13}$$

$$= \left[-\mathbb{E}_{y_{\text{data}}}\left(\frac{\partial^2\ell}{\partial\psi^a\,\partial\psi^b}\right)\right]_{\psi=\theta} = -\mathbb{E}_{y_{\text{data}}}\left(\frac{\partial^2\ell}{\partial\theta^a\,\partial\theta^b}\right) \tag{3.14}$$

where it was assumed that the integrand of the Kullback–Leibler divergence is regular enough such that the order of differentiation and integration may be exchanged. Further, one can confirm that if the order of differentiation can be exchanged, it is also true that the expectation value of the components of the score vanishes since

$$\mathbb{E}\left(\frac{\partial\ell}{\partial\theta^j}\right) = \mathbb{E}\left(\frac{1}{L}\,\frac{\partial L}{\partial\theta^j}\right) = \int_{\mathcal{D}} \mathrm{d}^N y_{\text{data}}\, \frac{\partial}{\partial\theta^j} L(y_{\text{data}} \,|\, \theta) = \frac{\partial}{\partial\theta^j} \underbrace{\int_{\mathcal{D}} \mathrm{d}^N y_{\text{data}}\, L(y_{\text{data}} \,|\, \theta)}_{=1} = 0. \tag{3.15}$$

By further application of $\frac{\partial}{\partial\theta^i}$ on both sides, it is found that

$$0 = \frac{\partial}{\partial\theta^i}\, \mathbb{E}_{y_{\text{data}}}\left(\frac{\partial\ell}{\partial\theta^j}\right) = \int \mathrm{d}^N y_{\text{data}}\, \frac{\partial}{\partial\theta^i}\left(L\,\frac{\partial\ell}{\partial\theta^j}\right) = \int \mathrm{d}^N y_{\text{data}} \left[\underbrace{\frac{\partial L}{\partial\theta^i}}_{L\frac{\partial\ell}{\partial\theta^i}}\,\frac{\partial\ell}{\partial\theta^j} + L\,\frac{\partial^2\ell}{\partial\theta^i\,\partial\theta^j}\right] \tag{3.16}$$

$$= \mathbb{E}\left(\frac{\partial\ell}{\partial\theta^i}\,\frac{\partial\ell}{\partial\theta^j}\right) + \mathbb{E}\left(\frac{\partial^2\ell}{\partial\theta^i\,\partial\theta^j}\right) \quad \Longrightarrow \quad \mathbb{E}_{y_{\text{data}}}\left(\frac{\partial\ell}{\partial\theta^i}\,\frac{\partial\ell}{\partial\theta^j}\right) = -\mathbb{E}_{y_{\text{data}}}\left(\frac{\partial^2\ell}{\partial\theta^i\,\partial\theta^j}\right) \tag{3.17}$$

wherefore the Fisher metric can also be expressed in terms of a product of first derivatives of the log-likelihood. This alternative form is typically more convenient when computing the Fisher metric and related objects such as the Christoffel symbols numerically.

On the other hand, the expression of the Fisher metric in terms of the expectation over the Hessian of ℓ indicates that the Fisher metric is a measure for how sharply the likelihood is peaked,

which means that the determinant of the Fisher metric is larger if the likelihood is more localised. Conversely, if the likelihood is expected to be very spread out, the determinant of the Fisher metric is small. For this reason, the Fisher metric is also generally referred to as the "Fisher information matrix" since it can be interpreted as measuring how much information a likelihood function encodes about each of the parameters in precisely the above sense.

As the name suggests, this information matrix was first introduced by R. Fisher around 1922, although it was only about 20 years later in 1945 when C. Rao first specifically exploited the fact that it exhibits the transformation behaviour of a $\binom{0}{2}$-tensor field in order to give it the interpretation of a Riemannian metric on a space of probability distributions (see [5, 23]). As mentioned in section 2.15, it was then proven by Čencov in 1981 that this is in fact the unique metric which remains invariant under congruent embeddings via Markov morphisms (see [16, 41, 79]).

It was shown by Amari that there exists a whole family of so-called α-connections which are compatible with the Fisher metric and whose connection coefficient functions can be parametrised using a constant $\alpha \in (0, 1)$ as

$$ {}^{(\alpha)}\Gamma_{kij} \coloneqq \mathbb{E}_{y_{\text{data}}}\left(\frac{\partial \ell}{\partial \theta^k} \frac{\partial^2 \ell}{\partial \theta^i \, \partial \theta^j} + \left(\frac{1-\alpha}{2} \right) \frac{\partial \ell}{\partial \theta^k} \frac{\partial \ell}{\partial \theta^i} \frac{\partial \ell}{\partial \theta^j} \right). \tag{3.18} $$

3.4.1 Geometry of Normal Error Distributions

For a dataset containing correlated observations whose errors are distributed according to a multivariate normal distribution where the log-likelihood ℓ can be expressed as in equation (3.7), the components of the Hessian of ℓ are given by

$$ -\frac{\partial^2 \ell}{\partial \theta^a \, \partial \theta^b} = \frac{\partial \zeta^i}{\partial \theta^a} \left(\Sigma^{-1}\right)_{ij} \frac{\partial \zeta^j}{\partial \theta^b} + \frac{\partial^2 \zeta^i}{\partial \theta^a \, \partial \theta^b} \left(\Sigma^{-1}\right)_{ij} \zeta^j \tag{3.19} $$

$$ = \frac{\partial h^i}{\partial \theta^a} \left(\Sigma^{-1}\right)_{ij} \frac{\partial h^j}{\partial \theta^b} + \frac{\partial^2 h^i}{\partial \theta^a \, \partial \theta^b} \left(\Sigma^{-1}\right)_{ij} \Big(h(\theta) - y_{\text{data}} \Big)^j \tag{3.20} $$

where it was assumed that the covariance matrix Σ does not depend on the parameters θ and the abbreviation $\zeta^i = \Big(y_{\text{data}} - h(\theta) \Big)^i$ was used. Subsequent application of the expectation value with respect to the observations y_{data} yields

$$ g_{ab}(\theta) = -\mathbb{E}_{y_{\text{data}}}\left(\frac{\partial^2 \ell}{\partial \theta^a \, \partial \theta^b} \right) \tag{3.21} $$

$$ = \int_{\mathcal{D}} \mathrm{d}^N y_{\text{data}} \, L(y_{\text{data}} \,|\, \theta) \left[\frac{\partial h^i}{\partial \theta^a} \left(\Sigma^{-1}\right)_{ij} \frac{\partial h^j}{\partial \theta^b} + \frac{\partial^2 h^i}{\partial \theta^a \, \partial \theta^b} \left(\Sigma^{-1}\right)_{ij} \Big(h(\theta) - y_{\text{data}} \Big)^j \right] \tag{3.22} $$

$$ = \frac{\partial h^i}{\partial \theta^a} \left(\Sigma^{-1}\right)_{ij} \frac{\partial h^j}{\partial \theta^b} + \frac{\partial^2 h^i}{\partial \theta^a \, \partial \theta^b} \left(\Sigma^{-1}\right)_{ij} \underbrace{\int_{\mathcal{D}} \mathrm{d}^N y_{\text{data}} \, L(y_{\text{data}} \,|\, \theta) \Big(h(\theta) - y_{\text{data}} \Big)^j}_{=0} \tag{3.23} $$

$$ = \frac{\partial h^i}{\partial \theta^a} \left(\Sigma^{-1}\right)_{ij} \frac{\partial h^j}{\partial \theta^b} \tag{3.24} $$

$$ \tag{3.25} $$

which is the final expression for the Fisher metric for observations with normal error distributions.

3.5 Central Objects of Information Geometry

This section provides a synopsis of the geometric embedding picture which is used by Transtrum et at. for example in [76–78, 83] for the statistical analysis of data whose errors are independently distributed according to a normal distribution. Ultimately, the aim is to extend this idea to general error distributions (see section 3.7).

As pointed out by Transtrum et al. in [78], the most important objects to consider in the information geometric modelling of data are the following:

1. the parameter manifold $\mathcal{M}$, which is the manifold of all admissible parameter configurations on which the model y_{model} can be evaluated,
2. the data space $\mathcal{D}$, which is a product space $\mathcal{D} \coloneqq \mathcal{Y}^N = \underbrace{\mathcal{Y} \times \ldots \times \mathcal{Y}}_{N \text{ times}}$ such that all N observations of a dataset constitute a single point $y_{\text{data}} \coloneqq (y_1, \ldots, y_N) \in \mathcal{D}$ and
3. the embedding map $h : \mathcal{M} \longrightarrow \mathcal{D}$ which connects the two spaces and is constructed from the model map y_{model}.

To be precise, the embedding map $h : \mathcal{M} \longrightarrow \mathcal{D}$ is defined as

$$h(\theta) \coloneqq \Big(y_{\text{model}}(x_1;\theta), \ldots, y_{\text{model}}(x_N;\theta)\Big) \equiv \bigotimes_{j=1}^{N} y_{\text{model}}(x_j;\theta) \in \mathcal{D} \tag{3.26}$$

and therefore encodes the collection of all predictions of the model y_{model} corresponding to the observation conditions $x_1, \ldots, x_N$ in the dataset. A graph illustrating the relationship between these three objects is shown in figure 25.

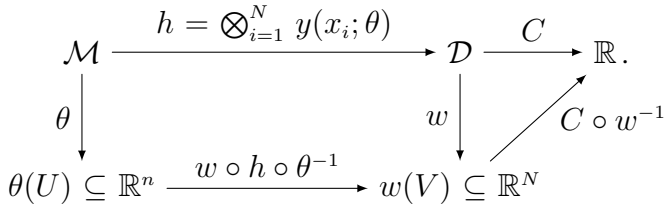


Figure 25: This graph depicts the embedding of the parameter manifold $\mathcal{M}$ into the data space $\mathcal{D}$ via the embedding map h and how this relationship can be expressed in coordinate charts (U, θ) and (V, w) on $\mathcal{M}$ and $\mathcal{D}$ respectively. Further, the function $C : \mathcal{D} \longrightarrow \mathbb{R}$ defined in equation (3.27) is used to judge how close the prediction $h(\theta_{\text{MLE}})$ of the model is to the observed data y_{data}.

The crucial insight about the embedding picture offered in [78] is the fact that for independent observations with normally distributed errors, the data space $\mathcal{D}$ constitutes a vector space. Immediately, this has a number of important consequences:

1. For models which depend linearly on the parameters $\theta \in \mathcal{M}$, the embedding map $h : \mathcal{M} \longrightarrow \mathcal{D}$ is also linear with respect to the parameters. Therefore, the vector space structure of $\mathcal{D}$ can be inherited to the parameter manifold $\mathcal{M}$, turning it into a vector space in its own right. It is then not hard to see that minimisation of the cost function C (i.e. the χ^2 test statistic) with respect to the parameters $\theta \in \mathcal{M}$ is equivalent to the minimisation of the metric distance between y_{data} and the collective prediction of the model $h(\theta) \in \mathcal{D}$. This recovers the well-known result that the ordinary least squares estimator is the best linear unbiased estimator (BLUE). This is also known as the Gauss–Markov theorem.

2. As suggested by the term "embedding", the map $h : \mathcal{M} \longrightarrow \mathcal{D}$ is required to be injective. If the number of parameters $n = \dim \mathcal{M}$ is larger than $N = \dim \mathcal{D}$ a map $h : \mathcal{M} \longrightarrow \mathcal{D}$ can no longer be injective and therefore such an embedding is no longer possible.

3. The Fisher metric $g_{\mathcal{M}}$ must correspond to the pull-back of a constant metric tensor on $\mathcal{D}$ along h for the embedding of $\mathcal{M}$ to be consistent with the ambient space $\mathcal{D}$.

4. Since $\mathcal{D}$ is a vector space it must be flat, i.e. its Ricci curvature $R = 0$.

It is observed in section 3.8.1 that the case of independent data with normally distributed errors is not the only example of an error distribution which induces a natural vector space structure on the data space. Consequently, the above observations are equally valid in those cases.

It should be noted that the definition of the data space $\mathcal{D}$ as adopted here differs slightly from the definition used by Transtrum et al. in the following way: in the context of this thesis, the data space $\mathcal{D}$ is just the product space of the observations and its coordinates are therefore also to be understood in the units in which the observations where made. On the other hand, Transtrum et al. adopt a pre-processing step in [78] where the data space is not the product space of the bare observations, but instead of the residuals $\big(y_i - y_{\text{model}}(x_i;\theta)\big)^2/\sigma_i^2$. As a result, what is referred to as the collective data point y_{data} here always corresponds to the origin in the space of residuals. However, since this particular form of pre-processing by computing the residuals is specific to quantities which are normally distributed and the goal is ultimately to generalise this formalism to other error distributions, it was chosen to omit this step here.

The cost function $C : \mathcal{D} \longrightarrow \mathbb{R}$ measures how close any point in the data space is to the point defined by collective observation data y_{data} and is thus subject to minimisation. A sensible candidate for this cost function is given by

$$C(\tilde{y}) = \frac{\chi^2(\tilde{y})}{2} = \frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i - \tilde{y}_i}{\sigma_i}\right)^2 \tag{3.27}$$

where $\chi^2(\tilde{y})$ denotes Pearson's χ^2 test statistic.[43] Comparing this to the log-likelihood, which can also be viewed as a function on the data space $\ell : \mathcal{D} \longrightarrow \mathbb{R}$, one recognises $\ell(y) = \text{const.} - C(y)$ which means that maximisation of the log-likelihood is also equivalent to minimisation of the cost function on $\mathcal{D}$.

[43] While some authors refer to C as a "cost function", others instead refer to it as the χ^2 distance.

Moreover, one may recognise that an appropriate metric function on $\mathcal{D}$ is given by

$$d_{\mathcal{D}}(p,q) = \sqrt{(p-q)^{\top}\,\Sigma^{-1}\,(p-q)} \tag{3.28}$$

where the inverse covariance matrix Σ^{-1} accounts for the uncertainties and in general the covariance in the measured data. In a slightly different context, this expression is also known as the Mahalanobis distance (see [21]). In other words, this metric function can be understood as being generated by an inner product on $\mathcal{D}$ which is induced via the inverse covariance matrix Σ^{-1} as described in equation (2.47). As a result, the cost function can also be written as

$$C(\tilde{y}) = \frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i-\tilde{y}_i}{\sigma_i}\right)^2 = \frac{1}{2}\Big(d_{\mathcal{D}}(y_{\text{data}},\tilde{y})\Big)^2 \tag{3.29}$$

since the definition of $d_{\mathcal{D}}$ recovers the χ^2 test statistic for diagonal covariance matrices Σ, i.e. independent measurements.

Further, using the metric $d_{\mathcal{D}}$ one can of course consider its induced metric balls, which correspond to concentric hyperellipses around $y_{\text{data}} \in \mathcal{D}$. That is, the metric balls correspond to the set of all points, which are separated from $y_{\text{data}} \in \mathcal{D}$ by less than a fixed metric distance of $r \in \mathbb{R}_0^+$ and can be expressed as $B_r(y_{\text{data}}) = \Big\{z \in \mathcal{D} \;\Big|\; d_{\mathcal{D}}(y_{\text{data}},z) \leq r\Big\}$. It is not hard to see that all points on the boundary of such a metric ball $B_r(y_{\text{data}})$ approximate the observed data equally well. The pull-backs of these metric-induced balls then correspond to the confidence regions in $\mathcal{M}$ (see section 4).

Given an embedding map h expressed in two charts $(U \subseteq \mathcal{M}, \theta)$ and $(V \subseteq \mathcal{D}, w)$, one can compute the push-forward map in components using equation (2.117) at $p \in \mathcal{M}$ by

$$(h_*)^j{}_a(p) = (\mathrm{d}w^j)_{h(p)}\left(h_*\left(\frac{\partial}{\partial\theta^a}\right)\right)_p = \left(\frac{\partial(w\circ h)^j}{\partial\theta^a}\right)_p \tag{3.30}$$

where the chain rule is executed in the last equality by acting on an arbitrary smooth function $f \in C^\infty(\mathcal{M})$ and observing that for any f one has

$$h_*\left(\frac{\partial}{\partial\theta^a}\right)(f) = \frac{\partial}{\partial\theta^a}(f\circ h) = \partial_a(f\circ h\circ\theta^{-1}) = \partial_a\Big((f\circ w^{-1})\circ(w\circ h\circ\theta^{-1})\Big) \tag{3.31}$$

$$= \partial_j(f\circ w^{-1})\cdot\partial_a(w^j\circ h\circ\theta^{-1}) = \frac{\partial(w^j\circ h)}{\partial\theta^a}\,\frac{\partial}{\partial w^j}\,f. \tag{3.32}$$

Having computed the derivative of the embedding map (i.e. its Jacobian) in components for a particular model once, it can subsequently be used to both push forward as well as pull back objects between the two manifolds. One can confirm that this embedding map h is indeed chosen consistently by verifying that the pull-back of the inner product (i.e. the metric tensor field) from $\mathcal{D}$ along h to $\mathcal{M}$ yields the same metric on the parameter manifold as one would obtain from approximating the Kullback–Leibler divergence through equation (2.206). Specifically, the pull-back

of the metric is performed by

$$(g_{\mathcal{M}})_{ab} = (h^* g_{\mathcal{D}})_{ab} = g_{\mathcal{D}}\left(h_*\left(\frac{\partial}{\partial\theta^a}\right), h_*\left(\frac{\partial}{\partial\theta^b}\right)\right) = (g_{\mathcal{D}})_{st}\,(h_*)^s{}_a\,(h_*)^t{}_b. \tag{3.33}$$

As was argued before, in the case that the errors are normally distributed, the metric on $\mathcal{D}$ should be taken as $(g_{\mathcal{D}})_{st} = (\Sigma^{-1})_{st}$ which yields the following metric on the parameter manifold:

$$(g_{\mathcal{M}})_{ab} = (h^* g_{\mathcal{D}})_{ab} = (h_*)^s{}_a\,(h_*)^t{}_b\,(\Sigma^{-1})_{st}. \tag{3.34}$$

By direct comparison, this can be seen to coincide exactly with the established formula for the Fisher metric in the case of a normal likelihood as stated in section 3.4.1. The above expression clearly illustrates that any kind of non-linear behaviour on $\mathcal{M}$ is a direct result of the model map and its parametrisation if the error distribution is normal and Σ is constant.

3.6 Parameter Identifiability

The topic of parameter identifiability is discussed in a series of papers by Transtrum, Machta, Sethna et al., which serve as the main source for this section (see e.g. [44, 68, 76–78, 83]).

Before starting the fitting process with the goal of ascertaining the optimal parameter configuration for a model to describe a given dataset, it should first be ensured that the problem of finding optimal parameters is actually well-posed. As outlined in [77, 83], there are two main criteria by which it can be judged whether a model map is well-conditioned in the sense that it is possible to find an optimal configuration using fitting algorithms. These criteria are referred to as structural identifiability and practical identifiability.

3.6.1 Structural Identifiability

To simplify the discussion, it is assumed that only error distributions whose derivative vanishes at a single point (i.e. the MLE) are considered, such that the same is true for the likelihood as a whole.

A model is defined to be (locally) structurally identifiable at a point $\theta \in \mathcal{M}$ if there exists a neighbourhood U around θ such that no other point $\psi \in U$ in this neighbourhood fulfils $y(x;\theta) = y(x;\psi)$ simultaneously for all $x \in \mathcal{X}$. Intuitively, one can take this to mean that a slight perturbation of the parameters θ in any direction has a quantifiable effect on the model prediction. This is illustrated by the following counterexample:

When considering $y(x;a,b) = a \cdot b \cdot x$ as a model function with parameters a and b, there are an infinite number of combinations $(a,b) \in \mathbb{R}^2$ which lead to an identical prediction. In other words, the parameters a and b are not structurally identifiable on their own—only the combination $m = a \cdot b$ is structurally identifiable. Their relationship can also be summarised as $b = m/a$.

Therefore, for every identifiable $m \in \mathbb{R}$ there is an equivalence class of parameter configurations which forms a curve in the two-dimensional parameter space. This implies that at every point

$\theta = (a, b) \in \mathcal{M}$ in the parameter space, there is a direction in which the likelihood does not change. Since the gradient of the likelihood must therefore vanish in this direction, the Fisher metric g is singular, i.e. $\det(g) = 0$.

Another example of a model where this argument applies analogously, is if two parameters are combined by an addition such as in $y(x; a, b) = (a + b) \cdot x$, leading to a relationship $b = m - a$ where again only $m \in \mathbb{R}$ is identifiable. For both examples, a visualisation of their associated unidentifiable curves on which every configuration leads the identical prediction is shown in figure 26. Intuitively, one may interpret the directions specified by the tangents of these curves at every point in the parameter manifold as uninformative directions.

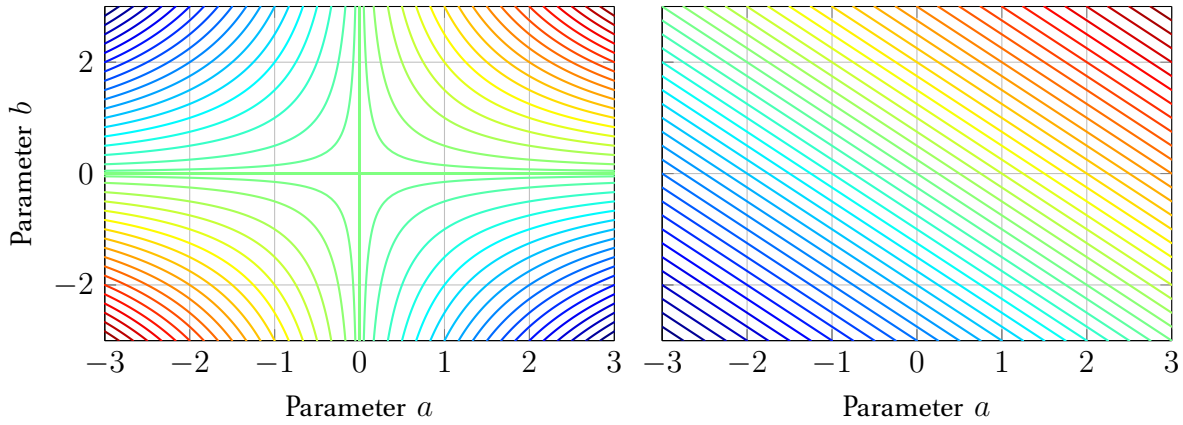


Figure 26: Plot of curves of unidentifiable parameter combinations for different values of the identifiable combination m. The left-hand side shows curves for a parametric relation of $b = m/a$, whereas the plot on right-hand side assumes a parametric relation of $b = m - a$. Essentially, every combination $(a, b) \in \mathbb{R}^2$ lies on such a curve.

Since tensors which vanish in one coordinate system must vanish in all coordinate systems, the determinant of the metric being zero is a manifold invariant.[44] As a result, models with a local structural unidentifiability should be considered ill-defined.

Although it was only argued above that singularity of the Fisher metric is a sufficient criterion for local structural unidentifiability, there actually exists a proof by T. Rothenberg in [62] to the effect that

$$\det\Big(g(\theta)\Big) \neq 0 \qquad \Longleftrightarrow \qquad \text{model is locally structurally identifiable at } \theta \in \mathcal{M}. \tag{3.35}$$

In essence, the proof consists of showing that if the metric g is singular at a point, one can always locally construct a curve along which the Fisher metric g is constant. On the other hand, Rothenberg shows that the Fisher metric g must be singular if the model is not locally structurally identifiable by use of the mean value theorem and the Bolzano–Weierstraß theorem. Apart from a few minor technical requirements, the proof only assumes that the log-likelihood ℓ is continuously differentiable and that the component functions of the resulting metric g are continuous functions

[44] This is due to the fact that the determinant of the metric is a tensor density of weight $w = +2$ which means that coordinate transformations can only result in multiplicative factors.

everywhere on $\mathcal{M}$. From the proof it is also clear that the total number of structurally unidentifiable combinations of parameters in a model is given by $\dim[\ker(g)]$, that is, by the dimension of the nullspace of the metric at a point.

As pointed out in [83], such structurally unidentifiable parameter combinations which cannot be determined by experiment are not only a problem in phenomenological science but are also frequently encountered throughout fundamental physics such as the charge and mass of particles, which can only be measured as a ratio in many physical theories.

In summary, the determinant of the Fisher metric offers a simple and convenient way of numerically detecting redundancies in the parametrisation of a given model, especially for complicated models which feature a large number of parameters. Yet, even if a model is locally structurally identifiable everywhere, it may still exhibit some undesirable properties with is shown by the further example:

Although the parameters $(a, b) \in (\mathbb{R}^+ \setminus \{0\}) \times (\mathbb{R}^+ \setminus \{0\})$ of the model $y(x; a, b) = e^{-ax} + e^{-bx}$ are not directly related by a simple equation, they are still symmetric under an exchange $(a, b) \mapsto (b, a)$. Geometrically, this corresponds to the existence of a line in $\mathcal{M}$ with respect to which the likelihood has a mirror symmetry. Careful calculation reveals that this model is locally structurally identifiable everywhere except on the $45°$ line defined by $a = b$ in the parameter space $\mathcal{M}$. While this may not impede the fitting process in itself provided that the initial configuration is not on the $45°$ line, it nonetheless complicates theoretical investigations of the model.

This motivates one to define the additional criterion of global structural identifiability: a model y is globally structurally identifiable if for any distinct pair $\theta, \psi \in \mathcal{M}$ with $\theta \neq \psi$ (i.e. not only in a small neighbourhood), there exist $x \in \mathcal{X}$ such that $y(x; \theta) \neq y(x; \psi)$. This means that the predictions of a model y are different for any mutually distinct parameter configurations.

In effect, global structural identifiability is equivalent to the requirement that the model should be injective as a map $y : \mathcal{M} \longrightarrow C^0(\mathcal{X}, \mathcal{Y})$ with respect to the parameters $\theta \in \mathcal{M}$, i.e. that the model y should satisfy

$$y(x; \theta) = y(x; \psi) \quad \forall x \in \mathcal{X} \qquad \Longrightarrow \qquad \theta = \psi. \tag{3.36}$$

Moreover, it is straightforward to see that for globally structurally identifiable models, the map $h : \mathcal{M} \longrightarrow \mathcal{D}$ defined in equation (3.26) is guaranteed to be an embedding, whereas for locally structurally identifiable models, it may only be an immersion (see e.g. figure 14). Thus, global structural identifiability implies local structural identifiability but the converse does not hold.

While it appears as though global structural identifiability is a very strong and limiting requirement, a locally structurally identifiable model can usually be made globally structurally identifiable by sufficiently restricting the model manifold $\mathcal{M}$. In this example, by simply adding the additional constraint that $a < b$ without loss of generality, the model becomes globally structurally identifiable while retaining the same diversity of predictions as before.

3.6.2 Practical Identifiability

Compared with structural identifiability, it is much harder to come up with a quantitative definition of practical identifiability. Generally, it should encapsulate the phenomenology that not all parameters are usually constrained equally by data. That is, the predictions of a model may be very sensitive with respect to some (combinations of) parameters but less so towards others.

On the one hand, the reason for this may lie in a suboptimal parametrisation of the model which can be very hard to recognise and rectify in practice. On the other hand, the amount of available data may be insufficient or it may be infeasible due to the nature of the experiment to collect observations $y \in \mathcal{Y}$ under observation conditions $x \in \mathcal{X}$ which would constrain the parameters more strongly.

One idea which has been put forward is that the eigenvalues of the Fisher metric g correspond to the sensitivity of the model with respect to the combination of parameters in the direction of the associated eigenvector of parameters. That is, eigenvectors associated with comparatively small eigenvalues of the Fisher metric should indicate uninformative directions whereas relatively large eigenvalues correspond to sensitive directions of the model. Said uninformative directions with extremely small eigenvalues are then referred to as practically unidentifiable directions of the model. Additionally, models whose eigenvalues span multiple orders of magnitude are termed "sloppy". An illustration of this terminology is provided in figure 27. In [44], it is observed that a sloppy eigenvalue spectrum appears to be universally shared by many different models from various disciplines of science and engineering, such as systems biology, machine learning, solid state physics and others.

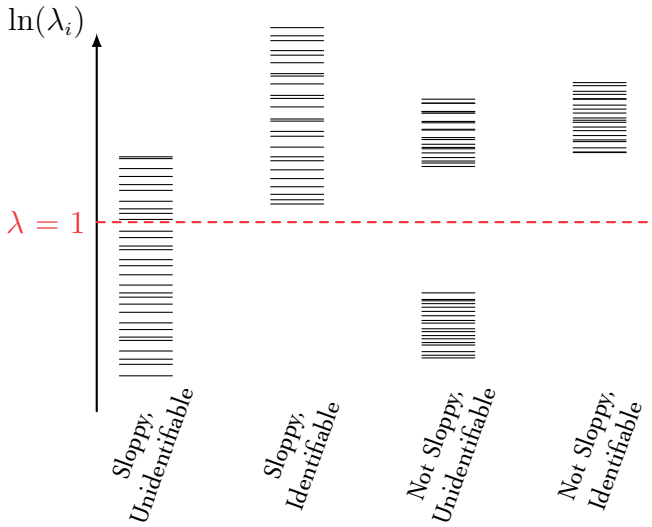


Figure 27: Illustration of the distinction between sloppiness and practical unidentifiability of model parameters in relationship to the "eigenvalues" λ_i of the Fisher metric g. Generally, a model is said to be sloppy (with respect to a certain choice of coordinates) whenever the eigenvalues span many orders of magnitude nearly uniformly on a logarithmic scale. On the other hand, a model is said to contain practically unidentifiable combinations of parameters if it has very small eigenvalues. This figure is reproduced from [83].

Clearly, this characterisation of practical unidentifiability suffers from several problems: first off, there are no quantitative criteria agreed upon by different authors to determine below which magnitude of eigenvalues a model should be considered practically unidentifiable or above what range of eigenvalues a model should be categorised as sloppy. More importantly, criteria of sloppiness and practical unidentifiability are crucially dependent on the parametrisation of the model, i.e. the choice of coordinates on $\mathcal{M}$. In other words, for any model which has been found to be sloppy and / or practically unidentifiable, there might in principle exist a coordinate transformation which renders it not sloppy and practically identifiable. The reason for this is that the eigenvalues of the metric are subject to change under reparametrisation.

Consequently, one can only infer whether the specific parametrisation that was chosen for the model is unsuitable from the eigenvalue analysis of the Fisher metric. However, this does not allow one to draw any conclusions about the fundamental properties of the model and the validity of its predictions itself, i.e. irrespective of parametrisation.

Since most models manifolds exhibit boundaries of some kind due to restrictions on the parameters, Transtrum et al. suggest the width of the parameter manifold $\mathcal{M}$ as measured by geodesic length as a coordinate invariant measure of practical parameter identifiability in a given direction. Using examples, they demonstrate that there appears to be good agreement between the magnitude of an eigenvalue and the geodesic width of the manifold in the associated eigendirection (see [78]). Finally, they conclude that this universally shared sloppiness of models means that the images of their parameter space under the respective embedding map $h(\mathcal{M}) \subseteq \mathcal{D}$ can generally be thought of and visualised as "hyper-ribbons". That is to say, the image of the model manifold typically features a hierarchy of progressively thinner widths due to its sloppy distribution of geodesic widths which means that it is very narrow in some directions but very long and drawn out in others.

In [76], Transtrum et al. propose an algorithm they term the "manifold boundary approximation method" or MBAM, which exploits the typically sloppy spectrum of models with large numbers of parameters to perform model reduction, i.e. to systematically eliminate the most uninformative parameter configurations of a model.

3.7 Proposed Extensions to the Information Geometric Picture

The following outlines a proposed extension to the embedding picture used by Transtrum et al. for example in [76–78, 83] which aims to abstract and generalise said embedding picture for a wider class of error distributions instead of being limited normally distributed errors.

So far, the data space $\mathcal{D} = \mathcal{Y}^N$ was just considered as constituting a set of ordered collections of real numbers which represent the possible observations and predictions. However, one can also think of every point in $p \in \mathcal{D}$ as corresponding to a probability distribution, whose location parameter (i.e. first moment) is given by the point. More specifically, each point point has an associated likelihood function which can be thought of as being centered on said collective observation.

If the location and covariance of the collective error distribution are independent then the likelihood

associated with each point $p \in \mathcal{D}$ only differs in its location. In particular, this is true if the error distribution is part of the location-scale family of probability distributions.

Being a space of probability distributions, one can then calculate the Kullback–Leibler divergences between any two points $p, q \in \mathcal{D}$ using the usual formula. Consequently, the Hessian of this divergence produces a metric on $\mathcal{D}$.

For normal likelihoods, one finds that the only component of the Hessian of the Kullback–Leibler divergence which ultimately contributes to the metric on $\mathcal{M}$ is the component associated with the second derivative with respect to the location parameter. In other words, the second derivative with respect to the location parameter of the Kullback–Leibler divergence between two multivariate normal distributions recovers exactly the inverse covariance matrix Σ^{-1}.

Analogously, the second derivative with respect to the location parameter of the Kullback–Leibler divergence between two (multivariate) Cauchy likelihoods can be calculated to find the appropriate factor which corresponds to the inverse covariance in the normal case (see section 3.8.1).

Thus, it is proposed that $\mathcal{D}$ should be regarded as a metric manifold in its own right. The geometric structures on $\mathcal{M}$ are then completely determined by the embedding map h. As a result it is in principle also possible to observe non-linear effects on the parameter manifold which are purely due to the error distribution of the data points instead of only due to the non-linearity of the embedding map (i.e. the modelling function).

3.7.1 The Initial Space $\mathcal{X}^N \times \mathcal{M}$

From the very beginning, the technical distinction between evaluating the model function $y_{\text{model}}(x; \theta)$ with respect to conditions $x_i \in \mathcal{X}$ and the parameters $\theta \in \mathcal{M}$ appears rather awkward in the formalism. Evidently, there must be some composite space on which the model map y_{model} is evaluated. This section focuses in detail on the technical implications of this initial space on which y_{model} is defined and demonstrates that its precise treatment is consistent with the more common informal treatment of model maps and embeddings.

In order to distinguish the map which connects the initial space[45] $\mathcal{X}^N \times \mathcal{M}$ to the data space $\mathcal{D}$ from what is usually referred to as the embedding $h : \mathcal{M} \longrightarrow \mathcal{D}$, the map $\tilde{h} : \mathcal{X}^N \times \mathcal{M} \longrightarrow \mathcal{D}$ is defined as

$$\tilde{h}(x_{\text{data}}; \theta) := \bigotimes_{i=1}^{N} y_{\text{model}}(x_i; \theta) = \Big(y_{\text{model}}(x_1; \theta), ..., y_{\text{model}}(x_N; \theta)\Big) \tag{3.37}$$

whereas the model map is already a map $y_{\text{model}} : \mathcal{X} \times \mathcal{M} \longrightarrow \mathcal{Y}$ and therefore remains unchanged.

As a set, the initial space is given by the Cartesian product $\mathcal{X}^N \times \mathcal{M}$. However, it makes sense to rig up this product space with as much structure as the constituent spaces naturally permit.

[45] The term "initial space" was chosen since this is where the model map is naturally evaluated. In hindsight, one might argue that this is a misnomer since $\mathcal{X}^N \times \mathcal{M}$ constitutes the intermediate space between $\mathcal{M}$ and $\mathcal{D}$.

That is, each individual $\mathcal{X}$ can be given a vector space structure, therefore also making the product $\mathcal{X}^N = \mathcal{X} \times \ldots \times \mathcal{X}$ into a vector space. Since any vector space is in particular also a smooth manifold, the initial space $\mathcal{X}^N \times \mathcal{M}$ naturally forms a smooth vector bundle.[46] That is, a fibre $F = \mathcal{X}^N$ which has a vector space structure is attached at each point $\theta \in \mathcal{M}$. Consequently, each point $\xi \in \mathcal{X}^N \times \mathcal{M}$ uniquely specifies a prediction associated with a set of conditions x_{data} and a parameter configuration $\theta \in \mathcal{M}$ for the given model, i.e. $\tilde{h}(\xi) \in \mathcal{D}$.

When attempting to identify the best fit point $\xi_{\text{MLE}} \in \mathcal{X}^N \times \mathcal{M}$ such that the model best describes a dataset, it is necessary to keep the measurement conditions $x_{\text{data}} \in \mathcal{X}^N$ fixed. To this end, one can provide a smooth section of this vector bundle, which is a map $Z : \mathcal{M} \longrightarrow \mathcal{X}^N \times \mathcal{M}$ on which the embedding h is then evaluated. In this case, the section that restricts the bundle to $(x_1, ..., x_N)$ is given with respect to the charts $(U \subseteq \mathcal{M}, \theta)$ and $(V \subseteq \mathcal{X}^N \times \mathcal{M}, \xi)$ by

$$(\xi \circ Z \circ \theta^{-1}) : \mathbb{R}^{\dim \mathcal{M}} \longrightarrow \mathbb{R}^{N \dim \mathcal{X} + \dim \mathcal{M}} \tag{3.38}$$

$$(\xi \circ Z \circ \theta^{-1})(\alpha_1, ..., \alpha_n) \coloneqq \Big(\overbrace{\underbrace{(x_1^1, ..., x_1^d)}_{=x_1}, x_2, ..., x_N}^{N \dim \mathcal{X}}, \overbrace{\alpha_1, ..., \alpha_n}^{\dim \mathcal{M}}\Big) \tag{3.39}$$

where the abbreviations $d = \dim \mathcal{X}$ and $n = \dim \mathcal{M}$ were used. From this, it is not hard to find that the push-forward Z_* is given in components by

$$(Z_*)^m{}_a = \frac{\partial \xi^m \circ Z}{\partial \theta^a} = \delta^m_{a+Nd} \quad \text{where} \quad m = 1, ..., Nd + n, \quad a = 1, ..., n \tag{3.40}$$

and can be expressed in matrix notation as

$$(Z_*)^m{}_a = \left(\frac{\mathbb{0}_{Nd \times n}}{\mathbb{1}_{n \times n}}\right)^m_a . \tag{3.41}$$

Thus, the pull-back of the metric along Z results in $Z^* g_{\mathcal{X}^N \times \mathcal{M}} = g_{\mathcal{M}}$ as expected.

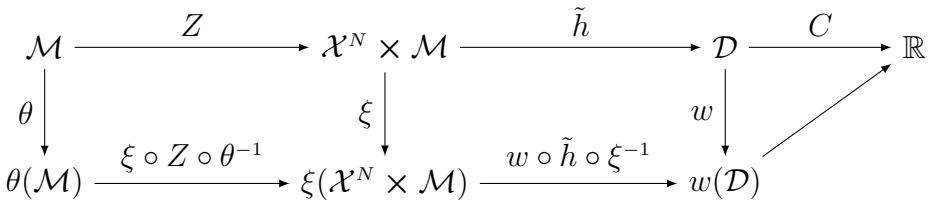


Figure 28: This graph depicts the maps between the spaces $\mathcal{M}$, $\mathcal{X}^N \times \mathcal{M}$ and $\mathcal{D}$ and their representations in coordinates. From this, one may also identify $h \equiv \tilde{h} \circ Z$.

Thus, the function to be maximised in order to obtain the best fit is $\ell \circ \tilde{h} \circ Z : \mathcal{M} \longrightarrow \mathbb{R}$. On

[46] The vector bundle $\mathcal{X}^N \times \mathcal{M} \xrightarrow{\pi} \mathcal{M}$ is equipped with the product topology and a smooth product atlas, such that $\mathcal{X}^N \times \mathcal{M}$ is again a smooth manifold.

top of this, one can also equip the vector space $\mathcal{X}^N$ with the standard inner product,[47] which is equivalent to choosing the euclidean flat metric on $\mathcal{X}^N$ such that the components of the metric tensor field in a chart are given by $g_{\mathcal{X}^N} = \mathrm{diag}(1, \dots, 1)$.

This allows for the definition of a product metric on all tangent spaces of $\mathcal{X}^N \times \mathcal{M}$ via

$$g_{\mathcal{X}^N \times \mathcal{M}}|_\xi : T_\xi(\mathcal{X}^N \times \mathcal{M}) \times T_\xi(\mathcal{X}^N \times \mathcal{M}) \longrightarrow \mathbb{R} \qquad \forall \xi \in \mathcal{X}^N \times \mathcal{M} \tag{3.42}$$

$$g_{\mathcal{X}^N \times \mathcal{M}} \coloneqq (\pi_{\mathcal{X}^N})^* \, g_{\mathcal{X}^N} + (\pi_{\mathcal{M}})^* \, g_{\mathcal{M}} \tag{3.43}$$

where $\pi_{\mathcal{X}^N}$ and $\pi_{\mathcal{M}}$ denote the the projection maps from $\mathcal{X}^N \times \mathcal{M}$ onto $\mathcal{X}^N$ and $\mathcal{M}$, respectively. This choice of metric on $\mathcal{X}^N \times \mathcal{M}$ can be considered "natural" in a sense due to the isomorphism of $T(\mathcal{X}^N \times \mathcal{M}) \cong_{\text{smooth}} (T\mathcal{X}^N) \times (T\mathcal{M})$.

In other words, the product metric is given by adding the pull-backs of the individual metrics along the respective projections. It is straightforward to check that this definition satisfies the requirements for a Riemannian metric. More precisely, the evaluation of the product metric is evaluated at any $\Big((a_{\mathcal{X}^N}, a_{\mathcal{M}}), (b_{\mathcal{X}^N}, b_{\mathcal{M}})\Big) = (A, B) \in T_\xi(\mathcal{X}^N \times \mathcal{M}) \times T_\xi(\mathcal{X}^N \times \mathcal{M})$ with $(x, \theta) = \xi \in \mathcal{X}^N \times \mathcal{M}$ is given by

$$(g_{\mathcal{X}^N \times \mathcal{M}})_\xi(A, B) = (g_{\mathcal{X}^N})_{\pi_{\mathcal{X}^N}(\xi)}\Big((\pi_{\mathcal{X}^N})_* \, A, (\pi_{\mathcal{X}^N})_* \, B\Big) + (g_{\mathcal{M}})_{\pi_{\mathcal{M}}(\xi)}\Big((\pi_{\mathcal{M}})_* \, A, (\pi_{\mathcal{M}})_* \, B\Big) \tag{3.44}$$

$$= (g_{\mathcal{X}^N})_x(a_{\mathcal{X}^N}, b_{\mathcal{X}^N}) + (g_{\mathcal{M}})_\theta(a_{\mathcal{M}}, b_{\mathcal{M}}) \tag{3.45}$$

where $(\pi_{\mathcal{M}*})^a{}_b = (\mathrm{d}(\pi_{\mathcal{M}}))^a{}_b = \frac{\partial(\theta \circ \pi)^a}{\partial \xi^b} = \delta^a_b$. In terms of components, one may picture this composite metric as

$$(g_{\mathcal{X}^N \times \mathcal{M}})_{ab} = \left(\begin{array}{c|c} g_{\mathcal{X}^N} & \mathbb{0}_{Nd \times n} \\ \hline \mathbb{0}_{n \times Nd} & g_{\mathcal{M}} \end{array}\right)_{ab} \tag{3.46}$$

where $\mathbb{0}$ denotes the zero matrix and indicates that there is no "mixing" between the components of the tangent spaces $T(\mathcal{X}^N)$ and $T\mathcal{M}$.

All in all, the above construction of the metric vector bundle $\mathcal{X}^N \times \mathcal{M}$ recovers the same metric on the initial space as proposed by Transtrum et al. in [78] for what is referred to as the "model graph" there. In said publication, it is argued that the addition of a parameter $\lambda \in \mathbb{R}^+$ into this composite metric as $(g_{\mathcal{X}^N \times \mathcal{M}})_{ab} = \lambda \cdot (\pi_{\mathcal{X}^N})^* \, g_{\mathcal{X}^N} + (\pi_{\mathcal{M}})^* \, g_{\mathcal{M}}$ allows for smooth interpolation between gradient descent and a Gauss-Newton method when this composite metric is used in the context of the Marquardt–Levenberg algorithm.

3.8 Alternative Geometries on the Data Space

This section studies examples of metric tensors induced by various error distributions on the data space $\mathcal{D}$ via the Kullback–Leibler distribution as detailed in section 3.7. Specifically, the focus is put

[47] If $\mathcal{X}$ is not a full vector space but only a convex cone, it can still be equipped with an inner product as shown in [61].

on error distributions for which the Kullback–Leibler divergence has a known analytical expression since the Fisher metric can be derived via the Hessian in a straightforward manner. Although the presented Kullback–Leibler divergences were published previously, it appears as though their information metrics have not been studied in detail so far. Of course, it is also possible to study more complicated error distributions where the Kullback–Leibler divergence is not known to have an analytical expression numerically.

3.8.1 Geometry of Cauchy Error Distributions

The calculation of the Kullback–Leibler divergence between two Cauchy distributions is rather involved but has been shown in [19] to yield

$$D_{\mathrm{KL}}\Big[p(x;\mu_1,s_1) : p(x;\mu_2,s_2)\Big] = \ln\left(\frac{(s_1+s_2)^2+(\mu_1-\mu_2)^2}{4\,s_1\,s_2}\right). \tag{3.47}$$

A second order expansion of the divergence between Cauchy error distributions (i.e. for a single observation $\mu = y$) is therefore given by

$$g_{11}(\mu,s) = \left[\frac{\partial^2}{\partial\tilde{\mu}\,\partial\tilde{\mu}}\,D_{\mathrm{KL}}\Big[p(x;\mu,s) : p(x;\tilde{\mu},\tilde{s})\Big]\right]_{\tilde{\mu}=\mu,\tilde{s}=s} = \frac{1}{2s^2} \tag{3.48}$$

which is very similar to the normal distribution, which yields σ^{-2}. As a result, for independent observations where each data point (x_i, y_i, s_i) is represented by a Cauchy distribution $\mathrm{Cauchy}(y_i, s_i)$, the data space metric tensor is given by $g_{\mathcal{D}} = \frac{1}{2}\mathrm{diag}(s_1^{-2},\ldots,s_N^{-2})$.[48] In particular, since $g_{\mathcal{D}}$ is constant, the data space is flat.

It might be tempting to conclude at this point, that the information contained in a data point whose error is Cauchy distributed is only half compared with the information content of a data point with whose error distribution is normal. Nevertheless, it is vital to keep in mind that the components of the Fisher metric are not manifold invariants. Rather, one should compare these metrics via the geodesics they induce, for example.

It is not hard to see that the shape of geodesics remains unaffected by this constant multiplicative rescaling. However, their length as determined by the metric-dependent length functional decreases by a factor of $1/\sqrt{2}$. On the other hand, by switching the error distribution which is used to model a dataset, one must also adapt the parameters for the new distribution to accurately reflect its properties. That is, since a Cauchy distribution is wider than a normal distribution, one would have to choose some $s < \sigma$ for their widths to become comparable. Because the variance of a Cauchy distribution is not finite (i.e. does not exist), one could instead compare their widths using the full width at half maximum (FWHM) which is given by $2\sqrt{2\ln(2)}\,\sigma$ for a normal distribution $N(\mu,\sigma)$ and $2s$ for a Cauchy distribution $\mathrm{Cauchy}(\mu,s)$.

Overall, it can be concluded that a Cauchy error distribution also induces a vector space structure on the data space $\mathcal{D}$. Since both the Cauchy and normal distributions are members of the family

[48] Note that the error density function is independent of x.

of student's t-distributions and represent the extremes of the spectrum in terms of the degrees of freedom,[49] one can conjecture that this is also the case for all t-distributions in between.

Moreover, since distributions in the student's t-family become more localised (i.e. their tails become leaner) with increasing number of degrees of freedom ν, one might further conjecture that the information content of any t-distribution must be larger or equal to that of the Cauchy distribution but smaller or equal to the information content of the normal distribution. That is, for any $\nu \in \mathbb{N}$ the Fisher information content associated with a change in the location parameter μ should be bounded by

$$\frac{1}{2s^2} \leq g_{11}(\mu, s, \nu) \leq \frac{1}{s^2} \tag{3.49}$$

where s denotes the scale parameter of the corresponding student's t-distribution. However, proving this bound explicitly for all values of ν is challenging. It may be prudent to resort to software-aided symbolic calculation.

3.8.2 Poisson Counting Errors for Normal Distribution

Whenever a measurement consists of counting a finite number of discrete events in a fixed amount of time, it can be assumed that the observations follow a Poisson distribution. Therefore, it is standard practice to estimate the errors in said measurements by the square root of the counts. That is, if it is observed that a phenomenon occurred exactly 100 times, the uncertainty in this measurement is estimated to be 10. A precise statistical analysis should therefore also take this direct relationship between the observed value and its uncertainty into account.

However, although the uncertainties are estimated using the Poisson distribution, it is typically still assumed for convenience that the error distribution is is normal, which in particular implies that the errors in the measurements are symmetric. Especially for large values of the observed counts, i.e. $\lambda \gg 1$, a normal distribution $\mathrm{Normal}(\mu, \sqrt{\lambda})$ provides a reasonable approximation to a Poisson distribution $\mathrm{Poisson}(\lambda)$. Nevertheless, this section aims to explore the implications of a direct coupling of the assumed uncertainties to their associated observations from the perspective of information content.

The Fisher information between two normal distributions whose moments are coupled as outlined above can be worked out as

$$g_{11}(\mu, \sigma) = \left[\frac{\partial^2}{\partial\tilde{\mu}\,\partial\tilde{\mu}} D_{\mathrm{KL}} \left[p\left(x; \mu, \sqrt{\mu}\right) : p\left(x; \tilde{\mu}, \sqrt{\tilde{\mu}}\right) \right] \right]_{\tilde{\mu}=\mu} \tag{3.50}$$

$$= \frac{1}{\mu} + \frac{1}{2\mu^2} = \frac{1}{2\mu^2}(2\mu + 1). \tag{3.51}$$

Since in this context $\mu = y_i = \sigma^2$, the information content associated with a data point $(x, y, \sigma = \sqrt{y})$ is given by $(y + 1/2)/y^2$ whereas a naïve treatment of the errors as independent of the y value would have resulted in an assumption of the information content as $\sigma^{-2} = 1/y$. In effect,

[49] Recall that for one degree of freedom $\nu = 1$, the student's t-distribution is exactly the Cauchy distribution, while the normal distribution is reached in the limit as $\nu \longrightarrow \infty$.

the term $1/(2y^2)$, which is always positive and therefore amounts to adding information content, would be missed by the naïve approach. For large values y, the effect of this correction term can be neglected, however, its contribution for small y cannot.

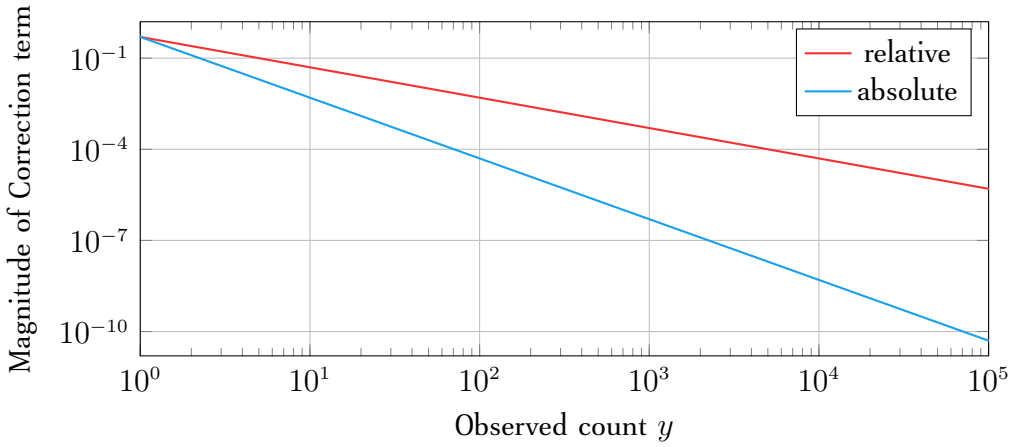


Figure 29: Information correction due to dependence of counting errors on the counted value y. This plot illustrates how this extra information content per data point decreases for large y. The absolute correction is given by $f_{\text{abs}}(y) = 1/(2y^2)$, whereas the relative correction is computed via $f_{\text{rel}}(y) = f_{\text{abs}}(y)/y^{-1} = 1/(2y)$.

Figure 29 provides a plot of the magnitude of the correction term. Although it appears to become inconsequential for values $y > 10^5$, its overall contribution when summed over all data points cannot be neglected when the number of data points N is on the order of the individual count numbers y.[50]

The most striking feature of this result is the fact that the metric $g_{\mathcal{D}}$ in the data space is no longer the constant inverse of a covariance matrix but instead a function of the position. Also, the geometric density factor $\sqrt{\det(g_{\mathcal{D}})}$ is no longer constant on the data space which indicates that the confidence regions are potentially no longer perfect hyperellipses.

Since the width of the error distribution is directly linked to the position y by construction, some kind of positional dependence is to be expected. Furthermore, the fact that the information content decreases for larger y is more than plausible as the width of the error distribution increases with y. While this dependence on position does not imply that the curvature on $\mathcal{D}$ is non-zero, it is sure to affect the metric on $\mathcal{M}$ after a pull-back along the embedding map h.

The metric can be expressed in components as

$$(g_{\mathcal{D}})_{ij} = \delta_{ij}\left(\frac{1}{y^i} + \frac{1}{2(y^i)^2}\right) \tag{3.52}$$

where no Einstein summation over the indices is implied due to the positioning of the indices on

[50] In the case that $y_i = N$ the relative contribution of the correction is exactly $\frac{1}{2}\sum_{i=1}^{N} y_i^{-1} = 1/2$.

the same level. Similarly, its inverse is given by

$$(g_{\mathcal{D}}^{-1})^{ki} = \delta^{ki}\left(\frac{1}{y^i} + \frac{1}{2(y^i)^2}\right)^{-1} = \delta^{ki}\,\frac{(y^i)^2}{y^i + 1/2}. \tag{3.53}$$

From this, one finds that the Christoffel symbols of the Levi-Civita connection are given by

$$\Gamma^k{}_{ja} = (g_{\mathcal{D}}^{-1})^{ki}\,\Gamma_{ija} = \underbrace{\delta^{ki}\,\delta_{ij}\,\delta_{ia}\,\delta_{ja}}_{\delta^k_j\,\delta^k_a}\left(-\frac{1}{2}\right)\left(\frac{y^i+1}{(y^i)^2 + y^i/2}\right) = \delta^k_j\,\delta^k_a\left(\frac{-(y^k+1)}{2(y^k)^2 + y^k}\right). \tag{3.54}$$

Therefore, especially in regions of the data space associated with small numbers of observed counts, the distance between points is no longer given by the euclidean distance between their coordinates. Instead, the geodesics connecting any points may be curved in coordinates.

When calculating the Riemann tensor from the given Christoffel symbols, one find $\mathrm{Riem}^k{}_{ikj} = 0$ for all $i, j, k, a = 1, ..., n$ everywhere. Thus, all other quantities which are built from contractions with the Riemann tensor, such as the Ricci tensor $\mathrm{Ric}_{ij} \coloneqq \mathrm{Riem}^k{}_{ikj} = 0$ and the Ricci scalar $R \coloneqq \mathrm{Ric}_{ij}\,(g_{\mathcal{D}}^{-1})^{ij} = 0$. Therefore, although the Christoffel symbols are non-trivial, the data space is still intrinsically flat.

This is not particularly surprising since although the metric $g_{\mathcal{D}}$ is a function of position, the underlying error distribution is still a normal distribution, which also yields a flat manifold in the usual case where the width is not connected to the position y.

3.8.3 Geometry of Gamma Error Distributions

The probability density of a gamma distribution is given by

$$p(x; k, \lambda) = \frac{x^{k-1}}{\lambda^k\,\Gamma(k)}\,\exp\!\left(-\frac{x}{\lambda}\right)\Theta(x) \tag{3.55}$$

with a shape parameter $k > 0$ and a scale parameter $\lambda > 0$. As these names suggest, the scale parameter λ affects the width of the distribution whereas the shape parameter k affects the location of the distribution although it doesn't directly correspond to the first moment. Instead, using its moment-generating function $M(t) = (1 - \lambda t)^{-k}$ one finds that the first moments of the gamma distribution are given by

$$\mathbb{E}(x) = k\,\lambda \qquad\qquad \mathbb{E}(x^2) = k\,\lambda^2. \tag{3.56}$$

The Kullback–Leibler divergence between two Gamma distributions has been shown in [11] to yield

$$D_{\mathrm{KL}}[\mathrm{Gamma}_1 : \mathrm{Gamma}_2] = (k_1 - k_2)\,\psi(k_1) + \ln\!\left(\frac{\Gamma(k_2)}{\Gamma(k_1)}\right) + k_2\,\frac{\ln(\lambda_2)}{\ln(\lambda_1)} + k_1\,\frac{\lambda_1 - \lambda_2}{\lambda_2}. \tag{3.57}$$

The Fisher information for the Gamma distribution subsequently works out to be

$$g_{ab}(k,\lambda) = \begin{pmatrix} \psi(k) & \lambda^{-1} \\ \lambda^{-1} & k\,\lambda^{-2} \end{pmatrix}. \tag{3.58}$$

where $\psi(k) = \Gamma'(k)/\Gamma(k)$ is the digamma function. A closer examination of its determinant reveals that this matrix is positive definite for values of $k \gtrsim 2.089$. The off-diagonal elements of the Fisher metric indicate, that the information contained in the two parameter k and λ is not independent, which is consistent with the moments stated in equation (3.56).

Given that the scale and shape of the gamma distribution are not mutually independent, it is not entirely clear how this can be accounted for using the geometry on $\mathcal{D}$.

4 Confidence in Manifolds

The goal of parameter inference is not only to find optimal parameter values such that a given model best describes observational data, but also to subsequently use this model to make predictions for the outcomes of future experiments. However, since any observation in the real world features statistical errors and noise, the chance that the model predictions coincide perfectly with future measurements is basically nil.

Therefore, it is absolutely necessary to quantify the uncertainties associated with each of the parameters such as to not render the predictions of the model meaningless. In turn, the uncertainties in the parameter configuration can then be used to estimate the uncertainty in the predictions of the model. This quantification of uncertainty associated with the parameters is achieved by establishing so-called confidence regions around the parameter configuration corresponding to the best fit.[51]

Roughly speaking, a confidence region of level $q \in [0,1]$ is a set of parameter configurations which contains the "true" parameter configuration with a probability of q, as estimated from the available data under the assumption that the given model is the correct description underlying the observed data. Moreover, a confidence region should contain only the most likely candidates for the "true" parameter configuration and is therefore usually a connected neighbourhood around the maximum likelihood estimate for globally structurally identifiable models.

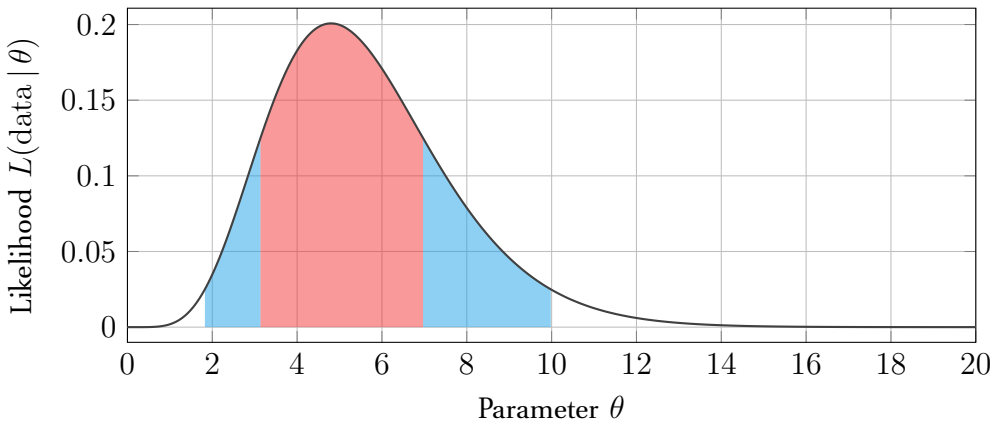


Figure 30: Illustration of an asymmetric likelihood function $L(\text{data}\,|\,\theta)$ which quantifies the probability of measuring the observed data from the given model with the parameter θ. The likelihood function can then be used to define likelihood-based confidence intervals as the set of parameter values with the highest likelihoods containing some fixed probability volume.

Concretely, for a model depending on a single parameter θ, i.e. $\dim \mathcal{M} = 1$, the confidence interval around the best fit point θ_{MLE} is constructed by

$$\mathbb{P}(a \leq \theta \leq b \,|\, \text{data}) = q \qquad \text{where} \qquad q \in [0,1] \quad \text{and} \quad \theta_{\text{MLE}} \in [a,b]. \tag{4.1}$$

[51] Although the term "confidence interval" technically only refers to models with a single parameter, this term is often used synonymously with "confidence region".

When generalising this to probability distributions that depend on more than one parameter, one has to make a distinction between individual confidence intervals and joint confidence regions. Constructing individual confidence intervals for each parameter $\theta^i \in \mathbb{R}$ separately presupposes that inferences about each of the parameters are made independently, which of course they are not. Furthermore, if one considers the product space of, say, four individual $95\,\%$ confidence intervals, the probability volume contained by this hypercube is not $95\,\%$ but typically more. Since they ultimately offer the only meaningful way of quantifying the uncertainty in parameter configurations, only simultaneous confidence regions are considered from this point onwards.

Instead, the shape of the boundary of a simultaneous confidence region depends both on the parametrisation of the model function and the error distribution. In particular, it is not hard to see that for normally distributed errors (i.e. normal likelihoods) and linearly parametrised model functions, the confidence regions must always form perfect hyperellipses centered around the maximum likelihood estimate θ_{MLE} on the model manifold $\mathcal{M}$.

This is because the iso-likelihood hypersurfaces in $\mathcal{D}$ already form hyperellipses around the combined data point $y_{\text{data}} \in \mathcal{D}$. Due to the linearity of the embedding map h with respect to the parameters, the push-forward h_* (i.e. the Jacobian) and similarly the pull-back h^* induced by h will be linear transformations. Since any linear distortion of an ellipsoid will again result in an ellipsoid, the confidence regions in $\mathcal{M}$ must therefore also be perfect hyperellipses. Moreover, the linear spacing between iso-likelihood surfaces in $\mathcal{D}$ for normal likelihoods also leads to a linear spacing between elliptic confidence regions under the pull-back to $\mathcal{M}$. The linearity of the embedding map h also effects a constant Fisher metric $g_{\mathcal{M}}$. Thus, it follows that the geometric density $\sqrt{\det(g_{\mathcal{M}})}$ must be constant for linearly parametrised models. However, in general the exact confidence region for non-linearly parametrised models can exhibit highly irregular and distorted shapes.

4.1 Defining Confidence Regions

Several definitions of simultaneous confidence regions have been proposed in the past, each of which comes with advantages and drawbacks. The most popular definitions of confidence regions are either based on the likelihood ratio test or the F-test. A summary of such methods can be found, for example, in [68]. A qualitative comparison between three different methods of establishing confidence regions for practical applications is given in [80].

To reiterate, a confidence region of level $q \in [0, 1]$ is a set of parameter configurations which is estimated to contain the "true" parameter configuration which underlies the observations with probability q. In addition, it is not just any set which contains the "true" parameter configuration with probability q: instead, it should only contain the parameter configurations which give the best possible descriptions of the data. That is, there should be no parameter configuration outside of a confidence region of some level q that describes the observed data as well or better than any parameter configuration inside the confidence region (e.g. has a higher likelihood).

What all widely used definitions of confidence regions have in common is that they exploit knowledge about the distribution of the test statistic to determine a threshold value below or

above which the test is said to reject a hypothesis with a confidence level q. Generally, the null hypothesis is that θ_{MLE} is the parameter configuration from which the observed data has been generated while the alternative hypothesis which is to be accepted or rejected is that the true parameter configuration is given by some other $\theta \in \mathcal{M}$. If the alternative hypothesis cannot be rejected according to the threshold value, the corresponding parameter configuration is said to belong to the confidence region of level q.

While it would be tempting to define confidence regions around the $\theta_{\text{MLE}} \in \mathcal{M}$ using the metric-induced open balls, for example via

$$q \overset{!}{=} \int_{\theta\left(B_r(p)\right)} \mathrm{d}^n\theta \sqrt{\left|\det\Big(g(\theta)\Big)\right|}\, L(\text{data}\,|\,\theta)\,, \tag{4.2}$$

it is not clear what the exact connection between geodesic distance and hypothesis tests such as the likelihood ratio is in general. For normal likelihoods, this indeed appears to coincide with the confidence regions established via the likelihood ratio test (see section 4.8, figure 43 and table 2). However, at this point in time, the necessary requirements for the likelihood are not known, e.g. whether it is possible to consistently adapt this for Cauchy likelihoods.

4.1.1 Confidence Regions Based on the Likelihood Ratio Test

The definition of confidence regions via the likelihood ratio test utilises Wilks' theorem (see section 4.4), which states that the log-likelihood difference is asymptotically distributed (i.e. in the limit that the number of data points $N \longrightarrow \infty$) according to χ^2_k. Thus, a confidence region of level q on the parameter manifold $\mathcal{M}$ may be defined as the set of parameter configurations given by

$$\mathcal{C}_q := \left\{\theta \in \mathcal{M} \;\middle|\; \ell(\theta_{\text{MLE}}) - \ell(\theta) \leq \frac{1}{2}\, F_k^{-1}(q)\right\} = \left\{\theta \in \mathcal{M} \;\middle|\; F_k\Big(2[\ell(\theta_{\text{MLE}}) - \ell(\theta)]\Big) \leq q\right\} \tag{4.3}$$

where F_k^{-1} denotes the inverse cumulative distribution function of the χ^2_k distribution (i.e. its quantile function) with k the number of parameters in which θ differs from θ_{MLE}. The application of F_k preserves the inequality which characterises $\mathcal{C}_q$ since the cumulative distribution of χ^2_k is a strictly monotonically increasing function.

The boundary of the confidence region $\partial\mathcal{C}_q$ is then obtained by considering only those parameter configurations, for which the defining inequality of $\mathcal{C}_q$ becomes an equality, i.e.

$$\partial\mathcal{C}_q = \left\{\theta \in \mathcal{M} \;\middle|\; \ell(\theta_{\text{MLE}}) - \ell(\theta) = \frac{1}{2}\, F_k^{-1}(q)\right\}. \tag{4.4}$$

Depending on the model function, the likelihood may have more than one local maximum, which can potentially result in topologically disconnected confidence regions. Although different authors disagree on whether topologically disconnected confidence regions should be allowed or excluded

by definition, it makes sense to require that any point $\theta \in \mathcal{M}$ be (path-)connected to θ_{MLE}.[52] This is because the additional condition of connectedness enables one to relate geodesic distance to the likelihood ratio and the Kullback–Leibler divergence in a consistent fashion (see section 4.8). Moreover, the existence of multiple local maxima in the likelihood can usually be traced back to a global structural unidentifiability of some kind in the model. As discussed in section 3.6, such global structural unidentifiabilities can be remedied by suitably restricting the parameter manifold $\mathcal{M}$.

The popularity of this approach stems at least partly from the Neyman–Pearson lemma (see [51]), which guarantees that the likelihood ratio is the most powerful test when comparing simple hypotheses. In addition, the likelihood ratio test is parametrisation-invariant and applicable in almost all practical settings.

However, since it can be cumbersome to evaluate accurately by hand for large datasets, Pearson himself proposed the χ^2-test as a convenient second order approximation to the likelihood ratio test in 1900 (see [56]). The χ^2-test statistic[53] for N independent measurements $\left\{(x_i, y_i, \sigma_i)\right\}$ with normally distributed errors is calculated by

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y_{\text{model}}(x_i; \theta)}{\sigma_i} \right)^2 . \tag{4.5}$$

Although the need for such approximations has arguably passed with the advent of modern digital computers, the use of this statistic has remained persistent throughout many branches of science. Moreover, it appears that many of its users are unaware that it was only ever intended as an approximation and that a much more accurate result is attainable with little more effort.[54]

4.1.2 Confidence Regions Based on the F-Test

Often, a dataset may be too small for Wilks' theorem, which relies on the assumption of an infinite amount of data, to yield an accurate approximation. An alternative to using the likelihood ratio test for the comparison of hypotheses is given by the so-called F-test. It is performed by evaluating the criterion

$$\frac{S(\theta)}{S(\theta_{\text{MLE}})} - 1 \leq \frac{k}{N-k} \, Q^{-1}_{k,N-k}(q) \quad \text{where} \quad S(\theta) \coloneqq \sum_{j=1}^{N} \left(\frac{y_j - y_{\text{model}}(x_j; \theta)}{\sigma_j} \right)^2 . \tag{4.6}$$

where $Q^{-1}_{k,N-k}$ denotes the quantile function of the Snedecor F-distribution (see section 2.13.5), which is also where this test gets its name. It is well-known that for $N \longrightarrow \infty$, the Snedecor

[52] That is, confidence regions $\mathcal{C}_q$ should be topologically connected for any confidence level $q \in [0, 1]$.

[53] Of course, the χ^2-test statistic defined in equation (4.5) is not to be confused with the χ^2_k probability distribution (see section 2.13.3).

[54] An extensive comparison of the higher order asymptotic behaviours of different hypothesis tests and approximations thereof can be found in [3].

F-distribution recovers the χ^2-distribution via

$$X \sim F(k, N-k) \qquad \Longrightarrow \qquad \left(\lim_{N\to\infty} kX\right) \sim \chi^2_k. \tag{4.7}$$

Apart from the fact that it is assumed that θ_{MLE} corresponds to the "true" parameter configuration, the F-test assumes that the errors are "spherically normal", i.e. jointly normally distributed.

Since the F-distribution is related to the t-distribution just like the χ^2-distribution is related to the normal distribution and a t-distribution is always more spread out compared with the normal distribution, it is not hard to see that the quantile function of the F-distribution will typically approach the quantile function of χ^2 from above, i.e. $k\,Q^{-1}_{k,N-k}(q) \gtrsim F^{-1}_k(q)$ for all $N \in \mathbb{N}$, except for very small values of q.

In principle, this relationship can therefore also be used to systematically overestimate the size of likelihood-based confidence regions by substituting the quantile function of the F-distribution for the quantile of the χ^2-distribution while retaining their relative shape.

4.2 Approximations of Confidence Regions

This section aims to emphasise the difference between exact confidence regions and approximations thereof. The most widely-used method for estimating the uncertainty associated with parameter configurations is given by the Cramér–Rao lower bound, which guarantees that the true covariance in the parameters is larger or equal to the inverse of the Fisher metric evaluated at the MLE.[55] In the event that the effect of the parameters is perfectly independent, i.e. if the covariance matrix is diagonal, one may conclude that a lower bound for the variances associated with the individual parameters is given by the diagonal entries of the inverse Fisher metric at the MLE.

However, the proof of the Cramér–Rao bound makes no statement about whether said lower bound is actually achieved in any specific case. Hence, this linear approximation not only fails to asymmetry in the sensitivity of the model with respect to a change in the parameters (even in the one-dimensional case of a model with a single parameter), the covariance ellipse typically further fails to provide a lower bound on the shape of the confidence region.

In anticipation of the results from section 4.3, the disparity between the exact (simultaneous) confidence regions and non-simultaneous confidence regions is highlighted both for a linearly parametrised model (see figure 31) as well as for a non-linearly parametrised model (see figure 32). Both of these illustrations are based on the "toy model" from section 5.1. As such, the non-linear reparametrisation is comparatively mild. In both cases, the individual confidence regions are depicted as dashed rectangles whose widths were obtained from the diagonal of the inverse Fisher metric at the MLE. As a result, the ellipse which is inscribed in the rectangles constitutes the Cramér–Rao lower bound for the covariance. It is evident from figures 31 and 32 that the lower

[55] Technically, the Cramér–Rao lower bound asserts that for unbiased estimators, the matrix given by $\Sigma_{\text{true}} - g^{-1}(\theta_{\text{MLE}})$ is positive definite where Σ_{true} denotes the true parameter covariance and $g^{-1}(\theta_{\text{MLE}})$ is the inverse of the Fisher metric as evaluated at the best fit (see e.g. [55]).

bound is typically not attained, even in the case of normal likelihoods and linearly parametrised models.

Thus, while it already misrepresents the true uncertainty in the parameters in these cases, its disparity compared with the exact confidence regions is even more pronounced for non-linearly parametrised models. Another example in which the disparity between the linear approximation of the uncertainty from the Cramér–Rao lower bound and the exact confidence region is even more substantial can be found in figure 55.

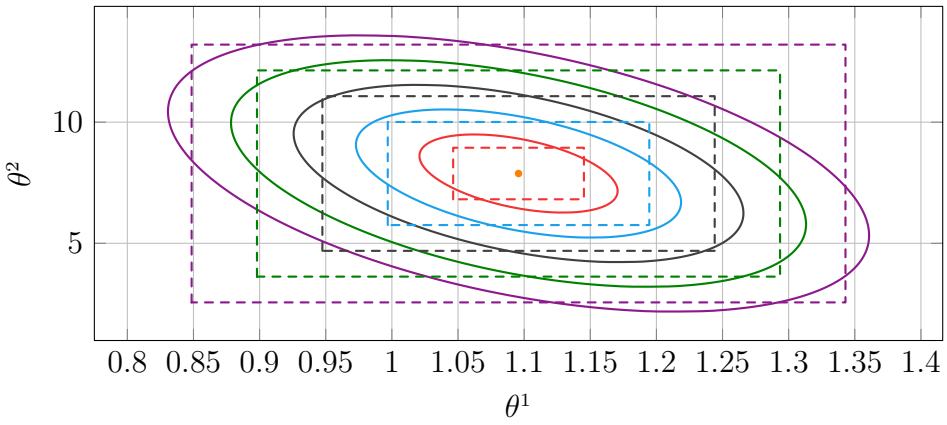


Figure 31: Comparison of joint confidence regions (ellipses) of confidence levels 1σ to 5σ against the corresponding individually constructed confidence intervals (rectangles) for a linearly parametrised toy model (see section 5.1). While the rectangular individual confidence intervals provide a decent approximation to the joint confidence regions for the confidence levels 2σ and 3σ, they go from underestimating the 1σ region to wildly overestimating the 4σ and 5σ regions. This shows that already in the case of linearly parametrised models, the naïve individual confidence interval estimates, as they are implemented in most commercial software, are an insufficient and potentially misleading way of quantifying the uncertainty in the parameters of a model.

As explored by Sellentin et al. in [69, 70], expanding the log-likelihood $\ell = \ln(L)$ in a Taylor series around the MLE as

$$-\ell(\psi) = -\sum_{n=0}^{\infty} \frac{1}{n!} \left[\frac{\partial}{\partial\theta^{a_1}} \dots \frac{\partial}{\partial\theta^{a_n}} \ln\big(p(\theta)\big) \right]_{\theta=\theta_0} (\theta-\theta_0)^{a_1} \dots \underbrace{(\theta-\theta_0)^{a_n}}_{=:\psi} \tag{4.8}$$

$$= -\ln\big(p(\theta_0)\big) - \underbrace{\left[\frac{\partial \ln(p)}{\partial\theta^a}\right]_{\theta=\theta_0}}_{=0} \psi^a - \frac{1}{2!} \underbrace{\left[\frac{\partial^2 \ln(p)}{\partial\theta^a\,\partial\theta^b}\right]_{\theta=\theta_0}}_{=:-F_{ab}} \psi^a\,\psi^b + \dots \tag{4.9}$$

$$= -\ln\big(p(\theta_0)\big) + \frac{1}{2!} F_{ab}\,\psi^a\,\psi^b + \frac{1}{3!} S_{abc}\,\psi^a\,\psi^b\,\psi^c + \frac{1}{4!} Q_{abcd}\,\psi^a\,\psi^b\,\psi^c\,\psi^d + \dots \tag{4.10}$$

allows one to achieve a more accurate approximation to the true likelihood, since asymmetries can be captured by the higher-rank tensors. The term DALI which is short for "derivative approximation for likelihoods" has been coined for this method. In particular, the higher-order corrections can be very useful to increase the efficiency of Monte Carlo methods which sample the likelihood.

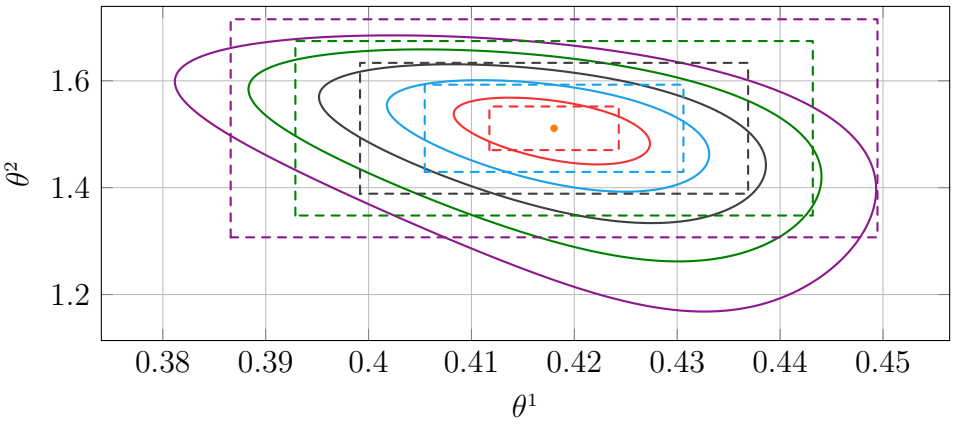


Figure 32: Comparison of joint confidence regions (round) of confidence levels 1σ to 5σ against the corresponding individually constructed confidence intervals (rectangles) for a non-linearly parametrised toy model (see section 5.1). The plot shows that the individual confidence intervals approximate the joint intervals only very roughly due to the inability of the rectangular shape to account for local changes in the geometric density and the resulting asymmetry in the joint intervals. Moreover, this approximation becomes increasingly inaccurate for larger confidence levels.

Compared with the integral manifold method from section 4.3, the DALI scheme only offers very granular control over the accuracy of the approximation via the number of terms which are included in the expansion of the log-likelihood. Since, it is not clear a priori how many terms must be considered in the expansion and how accurate the resulting DALI approximation is, there is no guarantee that the confidence regions obtained from this scheme faithfully represent the true parameter uncertainties.

4.3 Geometric Construction of Iso-Likelihood Surfaces

Given that the boundaries of confidence regions generally correspond to the level set of some function $f \in C^\infty(\mathcal{M})$ irrespective of the specific hypothesis test on which their definition is based, the method outlined in the following demonstrates how this can be exploited to find the exact boundary of confidence regions in a highly efficient manner. The general idea is to try to systematically find complete vector fields which are tangential to the level sets of f such that their integral curves or surfaces can be used to recover the entire level set.

If successful, this turns the problem of finding the boundary of a confidence region into a system of ordinary differential equations which can then be solved using modern numerical methods. The desired confidence level of the resulting boundary is specified by supplying a single point which is already known to lie on said boundary as an initial condition for the system of ODEs. This represents a significant reduction in computational effort, since the hypothesis test only needs to be evaluated on a one-dimensional line emanating from the maximum likelihood configuration $\theta_{\text{MLE}} \in \mathcal{M}$ to find such a point. Although this method was developed with the application of finding confidence regions in mind, it can be used to parametrise the level sets of any smooth function which satisfies the requirements discussed in section 4.3.1.

Given a scalar function $f \in C^\infty(\mathcal{M})$, its gradient is calculated using the exterior derivative, resulting in a covector field $\mathrm{d}f \in \Gamma(T^*\mathcal{M})$. Given such a covector field, one can try to find a vector field $X \in \Gamma(T\mathcal{M})$ such that in a chart (U, θ)

$$(\mathrm{d}f)(X) = X^j \frac{\partial f}{\partial \theta^j} \overset{!}{=} 0 \qquad \text{everywhere.} \tag{4.11}$$

In other words, the vector field X is annihilated by the gradient of f at every point. To call such a vector field orthogonal to $\mathrm{d}f$ would be somewhat unjustified, since no inner product is involved in this contraction. However, when identifying this covector field with its dual vector field via the sharp map as $\nabla f \coloneqq \sharp(\mathrm{d}f) = g^{-1}(\mathrm{d}f, \cdot) \in \Gamma(T\mathcal{M})$, considering ∇f as orthogonal to X with respect to the metric g gives the right intuition.

One may ponder the question of whether there are alternative principled ways of constructing vector fields which are tangential to the level sets of f, for example whether the construction should somehow account for geometric properties of $\mathcal{M}$ like curvature using the covariant derivative ∇_X. However, since both the covariant derivative ∇_X and also the Lie derivative $\mathcal{L}_X$ of a smooth function with respect to a vector field X by definition reduce to the same behaviour as the vector X acting on the function, one ends up with exactly the same criterion:[56]

$$\nabla_X f = \mathcal{L}_X f = Xf = (\mathrm{d}f)(X). \tag{4.12}$$

Intuitively, every one of these formulations aims to find a vector field along which the function f does not change in value. Disregarding the trivial vector field $X = 0$, a reasonable strategy for finding a general solution to equation (4.11) is to choose the components of X as

$$X^j = \alpha^j \prod_{i \neq j} \frac{\partial f}{\partial \theta^i} \qquad \text{for} \qquad \alpha^j \in \mathbb{R} : \sum_j \alpha^j = 0 \quad \text{and} \quad j = 1, \ldots, \dim \mathcal{M}. \tag{4.13}$$

By this method, every term in the Einstein summation of equation (4.11) is a product of all components of the gradient of f, and the annihilation of their sum is facilitated by the constant coefficients α^j which by construction sum to zero.[57] Inserting this form of X into equation (4.11), one finds

$$X^j \frac{\partial f}{\partial \theta^j} = \left(\sum_{j=1}^{\dim \mathcal{M}} \alpha^j \right) \underbrace{\prod_{i=1}^{\dim \mathcal{M}} \frac{\partial f}{\partial \theta^i}}_{=:B} \overset{!}{=} 0 \tag{4.14}$$

which, given that the product amounting to B is always non-zero for locally structurally identifiable models, vanishes only if $\sum_j \alpha^j = 0$. Moreover, one can see that for functions which are k times differentiable, the resulting vector field X will be $k - 1$ times differentiable. That is to say, X is

[56] Consequently, the vector field X can also be regarded as a formal Lie symmetry (see section 2.9.2) of the function f.

[57] There are of course other ways to construct vector fields that are solutions to equation (4.11), which may be simpler and less computationally expensive to evaluate, for instance, by dividing through by some components of $\mathrm{d}f$. However, depending on the particular components of $\mathrm{d}f$ this may have the unwanted effect of introducing points, lines or planes in $\mathcal{M}$ where the vector field diverges.

smooth if f is smooth.

Looking more closely at the condition $\sum_j \alpha^j \overset{!}{=} 0$, one can geometrically interpret this condition as a $(\dim \mathcal{M} - 1)$-dimensional hyperplane $\mathcal{H}$ in the real vector space $\mathbb{R}^{\dim \mathcal{M}}$ equipped with the standard inner product:

$$\mathcal{H} := \left\{ \vec{\alpha} \in \mathbb{R}^{\dim \mathcal{M}} \;\middle|\; \textstyle\sum_j \alpha^j = 0 \right\} = \left\{ \vec{\alpha} \in \mathbb{R}^{\dim \mathcal{M}} \;\middle|\; \vec{n} = (1, ..., 1)^\top,\ \vec{\alpha} \cdot \vec{n} = 0 \right\}. \tag{4.15}$$

This shows that any $\vec{\alpha}$ which is orthogonal to $\vec{n} = (1, ..., 1)^\top$ with respect to the standard inner product on $\mathbb{R}^{\dim \mathcal{M}}$ provides a solution to equation (4.14). Since, by definition, a vector space is n-dimensional if and only if it admits a set of n linearly independent basis vectors, it is clear that the hyperplane $\mathcal{H}$ must contain $\dim \mathcal{M} - 1$ vectors which are mutually orthogonal, as well as orthogonal to $\vec{n}$. In other words, $\dim \mathcal{H} = \dim \mathcal{M} - 1$.

Moreover, a basis of $\mathcal{H}$ can be constructed explicitly using the Gram–Schmidt algorithm, which generates an orthonormal basis from a set of linearly independent vectors. If the initial set consists of $\vec{n} = (1, ..., 1)^\top$ as well as $\dim \mathcal{M} - 1$ other linearly independent vectors, say, the standard basis vectors, then this algorithm produces an orthonormal basis which spans $\mathbb{R}^{\dim \mathcal{M}}$ and contains the vector $(\dim \mathcal{M})^{-1} (1, ..., 1)^\top$ as one of the basis elements. Clearly, the remaining basis elements must then constitute an orthonormal basis of $\mathcal{H}$.

As argued in section 2.3.2, any choice of inner product on some vector space V which is induced by a positive definite matrix is equivalent to choosing the standard inner product on V.[58] Since the Fisher metric g is Riemannian (i.e. positive definite), one can find isomorphisms $K_p : \mathbb{R}^{\dim \mathcal{M}} \longrightarrow T_p\mathcal{M}$ between $\mathbb{R}^{\dim \mathcal{M}}$ equipped with the standard inner product and any tangent space $(T_p\mathcal{M}, g_p)$ of the Riemannian manifold $(\mathcal{M}, g)$.

Accordingly, fixing an orthonormal basis of $\mathcal{H} \subset R^{\dim \mathcal{M}}$ also automatically determines an orthonormal basis of $\mathfrak{L}_p \subset T_p\mathcal{M}$ for each $p \in \mathcal{M}$ under the corresponding isomorphism $K_p : \mathbb{R}^{\dim \mathcal{M}} \longrightarrow T_p\mathcal{M}$. Because mapping a particular element $\vec{\alpha} \in \mathcal{H}$ under the isomorphism K_p for all $p \in \mathcal{M}$ specifies a unique element in every tangent space, this yields a vector field X on all of $\mathcal{M}$.

Specifically in the case where the log-likelihood function $f = \ell$ is considered, one can read off from equation (4.13) that the vector space isomorphism $K_p : \mathbb{R}^{\dim \mathcal{M}} \longrightarrow T_p\mathcal{M}$ must be given by

$$\left[K_p(\vec{\alpha})\right]^i = M^i{}_j\, \alpha^j = \left[\prod_{k=1}^{n} \frac{\partial \ell}{\partial \theta^k}\right] \operatorname{diag}\left(\left(\frac{\partial \ell}{\partial \theta^1}\right)^{-1}, ..., \left(\frac{\partial \ell}{\partial \theta^n}\right)^{-1}\right)^i{}_j\, \alpha^j \tag{4.16}$$

where $n = \dim \mathcal{M}$. As previously mentioned, the components of the score are non-zero for locally structurally identifiable models, wherefore the linear map K_p is invertible for every $p \in \mathcal{M}$ and X vanishes almost nowhere, i.e. only at extremal points of ℓ. To reiterate, this also means that the regularity of vector fields $X = K(\vec{\alpha})$ depends entirely on the regularity of the log-likelihood ℓ: if ℓ is k times continuously differentiable, then X is $k - 1$ times continuously differentiable.

[58] That is to say, there are only as many inequivalent choices of inner product as there are non-degenerate signatures.

The integral curves of X will then trace out level sets I_c of the log-likelihood ℓ defined by

$$I_c \coloneqq \operatorname{preim}_\ell(c) = \left\{\theta \in \mathcal{M} \;\middle|\; \ell(\theta) = c\right\} \tag{4.17}$$

given an initial condition in the form of a starting point which already lies on the desired level set. The defining equation for an integral curve γ to a vector field X is given by

$$X_{\gamma(t)} \overset{!}{=} \dot{\gamma}(t) \tag{4.18}$$

which enforces that the tangent vectors $\dot{\gamma}$ to the curve γ coincide with the vector field X at every point through which the curve passes. In a suitable chart (U, θ) where $\gamma(t) \in U$, this condition translates to

$$\left(X_{\gamma(t)}\right)^j \overset{!}{=} \left(\theta^j \circ \gamma\right)'(t) \tag{4.19}$$

which is a set of ordinary differential equations that is guaranteed to have a unique solution (at least locally) by virtue of the Picard–Lindelöf theorem, given appropriate initial conditions. More generally, the existence of integral surfaces or integral manifolds is characterised by the Frobenius theorem (see section 4.3.3), whose requirement that a set of vector fields should commute is trivially fulfilled in the one-dimensional case, where said set only consists of one vector field.

4.3.1 The Lie Subalgebra of Likelihood-Annihilating Vector Fields

The Frobenius theorem guarantees that the span of a set of vector fields $X_1, \ldots, X_k \in \Gamma(T\mathcal{M})$ generates a unique family of integral manifolds if said span of vector fields constitutes a closed Lie subalgebra of $\Gamma(T\mathcal{M})$ (see section 4.3.3). If this family of integral manifolds indeed exists, it is also guaranteed that it foliates $\mathcal{M}$. Therefore, this section aims to investigate whether vector fields of the form given in equation (4.13) constitute a closed Lie algebra. Specifically, the set of smooth vector fields of this form will be denoted by

$$\mathfrak{L} \coloneqq \left\{K(\vec{\alpha}) \in \Gamma(T\mathcal{M}) \;\middle|\; \vec{\alpha} \in \mathcal{H}\right\} \tag{4.20}$$

from here on out, where $K(\vec{\alpha})$ denotes the collection of $K_p(\vec{\alpha})$ for all $p \in \mathcal{M}$. By definition, one therefore has $\mathcal{L}_X\, \ell = 0$ for all $X \in \mathfrak{L}$, meaning that any element of $\mathfrak{L}$ annihilates the log-likelihood.

It is not difficult to come up with special examples where the level sets of the log-likelihood indeed foliate the parameter manifold $\mathcal{M}$. In those cases, Frobenius' theorem conversely states that the existence of a foliation implies the existence of a Lie subalgebra from which it can be generated. Therefore, it is more than plausible from the outset that $\mathfrak{L}$ forms a Lie subalgebra of the Lie algebra of all smooth vector fields if sufficient restrictions are placed on the model function and the log-likelihood (via the error distribution). The proof outlined in this section highlights that the necessary restrictions consist of the global structural identifiability of the model on the one hand and twice-continuous differentiability of the log-likelihood ℓ.

It is not hard to see that the set $\mathfrak{L}$ must be smaller than the set $\left\{X \in \Gamma(T\mathcal{M}) \;\middle|\; \mathcal{L}_X\, \ell \equiv X\ell = 0\right\}$

since not all vector fields which annihilate ℓ are necessarily of the form given in equation (4.13). That is, if $X \in \mathfrak{L}$, then any other smooth vector field Y which is related to X by a smooth function provides another valid solution to equation (4.11), i.e.

$$\forall X \in \mathfrak{L} : \forall f \in C^\infty(\mathcal{M}) : \qquad Y = fX \qquad \Longrightarrow \qquad \mathcal{L}_Y\, \ell = 0 \tag{4.21}$$

while generally $Y \notin \mathfrak{L}$. Using the vector space isomorphism $K_p : \mathbb{R}^{\dim \mathcal{M}} \longrightarrow T_p\mathcal{M}$, it immediately follows that $(\mathfrak{L}, +, \cdot\,)$ forms a (finite-dimensional) $\mathbb{R}$-vector subspace of $\Big(\Gamma(T\mathcal{M}), +, \cdot\,\Big)$ since $\mathcal{H}$ is an $\mathbb{R}$-vector subspace of $\mathbb{R}^{\dim \mathcal{M}}$. Thus, it only remains to be shown that $\mathfrak{L}$ is closed with respect to the Lie bracket, i.e. that $[X, Y] \in \mathfrak{L}$ for all $X, Y \in \mathfrak{L}$. To show this, it is again convenient to make use of the isomorphism K_p. The vector space $\mathbb{R}^{\dim \mathcal{M}}$ can be equipped with a Lie bracket $[\![\,\cdot\,,\,\cdot\,]\!] : \mathbb{R}^{\dim \mathcal{M}} \times \mathbb{R}^{\dim \mathcal{M}} \longrightarrow \mathbb{R}^{\dim \mathcal{M}}$ in such a way that it is compatible with the Lie bracket of smooth vector fields in the sense

$$K_p\Big([\![\vec{\alpha}, \vec{\beta}]\!]\Big) \stackrel{!}{=} \Big[K_p(\vec{\alpha}), K_p(\vec{\beta})\Big]. \tag{4.22}$$

Clearly, this condition is satisfied by just using K_p to define the bracket as

$$[\![\vec{\alpha}, \vec{\beta}]\!] \coloneqq K_p^{-1}\Big(\Big[K_p(\vec{\alpha}), K_p(\vec{\beta})\Big]\Big) \tag{4.23}$$

since K_p is invertible. From this definition, it follows that $\mathcal{H}$ and $\mathfrak{L}$ must be isomorphic as Lie algebras provided that they are both closed under their respective Lie brackets, which can be summarised as

$$\Big(\mathcal{H}, [\![\,\cdot\,,\,\cdot\,]\!]\Big) \cong_{\text{Lie alg.}} \Big(\mathfrak{L}, [\,\cdot\,,\,\cdot\,]\Big) \tag{4.24}$$

because the isomorphism K_p is valid at every point $p \in \mathcal{M}$. The problem of proving that $\mathfrak{L}$ is closed with respect to the Lie bracket $[\,\cdot\,,\,\cdot\,]$ is thus simplified to showing that $\mathcal{H}$ is closed in $\mathbb{R}^{\dim \mathcal{M}}$ with respect to the Lie bracket $[\![\,\cdot\,,\,\cdot\,]\!]$.

For any smooth vector fields $X, Y \in \Gamma(T\mathcal{M})$ one can express the Lie bracket in components as

$$[X, Y]f = X(Yf) - Y(Xf) = X^i \frac{\partial}{\partial \theta^i}\left(Y^j \frac{\partial f}{\partial \theta^j}\right) - Y^i \frac{\partial}{\partial \theta^i}\left(X^j \frac{\partial f}{\partial \theta^j}\right) \tag{4.25}$$

$$= X^i \left(\frac{\partial Y^j}{\partial \theta^i}\frac{\partial f}{\partial \theta^j} + Y^j \frac{\partial^2 f}{\partial \theta^i\, \partial \theta^j}\right) - Y^i \left(\frac{\partial X^j}{\partial \theta^i}\frac{\partial f}{\partial \theta^j} + X^j \frac{\partial^2 f}{\partial \theta^i\, \partial \theta^j}\right) \tag{4.26}$$

$$= \underbrace{\left(X^i \frac{\partial Y^j}{\partial \theta^i} - Y^i \frac{\partial X^j}{\partial \theta^i}\right)}_{=[X,Y]^j} \frac{\partial f}{\partial \theta^j} + \underbrace{\left(X^i\, Y^j - Y^i\, X^j\right) \frac{\partial^2 f}{\partial \theta^i\, \partial \theta^j}}_{=0} \tag{4.27}$$

where the last term vanishes due to the contraction of a symmetric with an antisymmetric quantity. By representing the linear transformation K_p via $\Big(K_p(\vec{\alpha})\Big)^i = M^i{}_j\, \alpha^j$, one can compute

$$\left(K_p^{-1}\Big(\Big[K_p(\vec{\alpha}), K_p(\vec{\beta})\Big]\Big)\right)^i = (M^{-1})^i{}_j \left(\underbrace{M^b{}_a\, \alpha^a}_{[K_p(\vec{\alpha})]^b}\, \partial_b\, \underbrace{M^j{}_c\, \beta^c}_{[K_p(\vec{\beta})]^j} - M^b{}_a\, \beta^a\, \partial_b\, M^j{}_c\, \alpha^c\right) \tag{4.28}$$

$$= (\alpha^a\,\beta^c - \alpha^c\,\beta^a)\,(M^{-1})^i{}_j\,M^b{}_a\,\partial_b\,M^j{}_c\,. \tag{4.29}$$

Further, one has

$$0 = \frac{\partial}{\partial\theta^b}\big(\delta^i_c\big) = \frac{\partial}{\partial\theta^b}\big((M^{-1})^i{}_j\,M^j{}_c\big) = \frac{\partial(M^{-1})^i{}_j}{\partial\theta^b}\,M^j{}_c + (M^{-1})^i{}_j\,\frac{\partial M^j{}_c}{\partial\theta^b} \tag{4.30}$$

from which it immediately follows that $(M^{-1})^i{}_j\,\partial_b\,M^j{}_c = -M^j{}_c\,\partial_b\,(M^{-1})^i{}_j$, i.e. the derivative can be shifted from the matrix M onto its inverse M^{-1} at the cost of a negative sign.

$$[\![\vec\alpha,\vec\beta]\!]^i = \left[K_p^{-1}\Big(\big[K_p(\vec\alpha),K_p(\vec\beta)\big]\Big)\right]^i = (\alpha^a\,\beta^c - \alpha^c\,\beta^a)\,(M^{-1})^i{}_j\,M^b{}_a\,\partial_b\,M^j{}_c \tag{4.31}$$

$$= -(\alpha^a\,\beta^c - \alpha^c\,\beta^a)\,M^j{}_c\,M^b{}_a\,\partial_b\,(M^{-1})^i{}_j \tag{4.32}$$

The partial derivatives of the coefficient functions of M^{-1} can be worked out as

$$\frac{\partial(M^{-1})^i{}_j}{\partial\theta^b} = \frac{\partial}{\partial\theta^b}\left[\frac{1}{B}\,\mathrm{diag}\left(\left(\frac{\partial\ell}{\partial\theta^1}\right),\ldots,\left(\frac{\partial\ell}{\partial\theta^n}\right)\right)^i{}_j\right] = \frac{\partial}{\partial\theta^b}\left[\frac{1}{B}\,\delta^i_j\,\frac{\partial\ell}{\partial\theta^j}\right] \tag{4.33}$$

$$= -\frac{1}{B^2}\,\frac{\partial B}{\partial\theta^b}\,\delta^i_j\,\frac{\partial\ell}{\partial\theta^j} + \frac{1}{B}\,\delta^i_j\,\frac{\partial^2\ell}{\partial\theta^b\,\partial\theta^j} = \frac{1}{B}\,\delta^i_j\left(\frac{\partial^2\ell}{\partial\theta^b\,\partial\theta^j} - \frac{\partial\ell}{\partial\theta^j}\,\frac{\partial\ln(B)}{\partial\theta^b}\right). \tag{4.34}$$

Reinserting this expression for the partial derivatives of M^{-1} yields

$$[\![\vec\alpha,\vec\beta]\!]^i = -(\alpha^a\,\beta^c - \alpha^c\,\beta^a)\,M^j{}_c\,M^b{}_a\,\partial_b\,(M^{-1})^i{}_j \tag{4.35}$$

$$= -(\alpha^a\,\beta^c - \alpha^c\,\beta^a)\,M^j{}_c\,M^b{}_a\,\frac{1}{B}\,\delta^i_j\left(\frac{\partial^2\ell}{\partial\theta^b\,\partial\theta^j} - \frac{\partial\ell}{\partial\theta^j}\,\frac{\partial\ln(B)}{\partial\theta^b}\right) \tag{4.36}$$

$$= -2\underbrace{\Big(K_p(\vec\alpha)\Big)^{[b}\Big(K_p(\vec\beta)\Big)^{j]}}_{\text{antisymm.}}\,\frac{1}{B}\,\delta^i_j\left(\underbrace{\frac{\partial^2\ell}{\partial\theta^b\,\partial\theta^j}}_{\text{symm.}} - \frac{\partial\ell}{\partial\theta^j}\,\frac{\partial\ln(B)}{\partial\theta^b}\right) \tag{4.37}$$

$$= 2\Big(K_p(\vec\alpha)\Big)^{[b}\Big(K_p(\vec\beta)\Big)^{j]}\,\frac{1}{B}\,\delta^i_j\,\frac{\partial\ell}{\partial\theta^j}\,\frac{\partial\ln(B)}{\partial\theta^b} \tag{4.38}$$

where the sum over the index j is inhibited by the Kronecker symbol δ^i_j which restricts the sum to the term corresponding to the open index i. Additionally, the expressions in equations (4.37) and (4.38) employ the commonly used antisymmetrisation bracket notation. Since the above expression is an element of the vector space $\mathbb{R}^{\dim\mathcal{M}}$, it remains to be shown that $\vec n\cdot[\![\vec\alpha,\vec\beta]\!] = 0$ in order to guarantee that $[\![\vec\alpha,\vec\beta]\!]\in\mathcal{H}$ which then concludes the proof that $\mathfrak{L}$ is a Lie subalgebra of $\Big(\Gamma(T\mathcal{M}),+,\cdot,[\,\cdot\,,\,\cdot\,]\Big)$. Finally, one obtains

$$\vec n\cdot[\![\vec\alpha,\vec\beta]\!] = \sum_{i=1}^{\dim\mathcal{M}}(1,\ldots,1)^i\cdot[\![\vec\alpha,\vec\beta]\!]^i = \sum_{i=1}^{\dim\mathcal{M}}[\![\vec\alpha,\vec\beta]\!]^i \tag{4.39}$$

$$= \sum_{i=1}^{\dim\mathcal{M}}2\Big(K_p(\vec\alpha)\Big)^{[b}\Big(K_p(\vec\beta)\Big)^{j]}\,\frac{1}{B}\,\delta^i_j\,\frac{\partial\ell}{\partial\theta^j}\,\frac{\partial\ln(B)}{\partial\theta^b} \tag{4.40}$$

$$= \frac{1}{B}\frac{\partial \ln(B)}{\partial \theta^b}\left[\left(K_p(\vec{\alpha})\right)^b \sum_{i=1}^{\dim \mathcal{M}} \delta^i_j \left(K_p(\vec{\beta})\right)^j \frac{\partial \ell}{\partial \theta^j} - \left(K_p(\vec{\beta})\right)^b \sum_{i=1}^{\dim \mathcal{M}} \delta^i_j \left(K_p(\vec{\alpha})\right)^j \frac{\partial \ell}{\partial \theta^j}\right] \tag{4.41}$$

$$= \frac{1}{B}\frac{\partial \ln(B)}{\partial \theta^b}\left[\left(K_p(\vec{\alpha})\right)^b \underbrace{\sum_{j=1}^{\dim \mathcal{M}} \left(K_p(\vec{\beta})\right)^j \frac{\partial \ell}{\partial \theta^j}}_{=0} - \left(K_p(\vec{\beta})\right)^b \underbrace{\sum_{j=1}^{\dim \mathcal{M}} \left(K_p(\vec{\alpha})\right)^j \frac{\partial \ell}{\partial \theta^j}}_{=0}\right] = 0 \tag{4.42}$$

where the summation over j is now executed without obstruction. This causes the expression to vanish due to the contraction of the components of the vector field $K_p(\vec{\alpha})$ with the derivatives of the log-likelihood which vanishes by construction for any $\vec{\alpha} \in \mathcal{H}$. Thus, the new element $\vec{\nu} = [\![\vec{\alpha}, \vec{\beta}]\!]$ must be in $\mathcal{H}$.

Since vector fields constructed via equation (4.13) evidently form a $(\dim \mathcal{M} - 1)$-dimensional Lie subalgebra of the infinite-dimensional Lie algebra of smooth vector fields, Frobenius' theorem guarantees that integral manifolds of this subalgebra always exist. Moreover, the outlined proof identifies the sufficient differentiability of ℓ as well as the global structural identifiability of the model as the key criteria[59] for the guaranteed existence of confidence regions.

Therefore, in the case of higher-dimensional confidence boundaries such as surfaces or manifolds in general, one can use the flows with respect to a basis of the Lie algebra $\mathfrak{L}$ to reach any point on a given confidence boundary starting from any other. Intuitively, this can also be imagined as meshing the confidence boundary using families of integral curves whose tangent vectors collectively form a $(\dim \mathcal{M} - 1)$-dimensional linear subspace of the tangent space $T_p\mathcal{M}$ at every point p on the confidence boundary.

Once a basis of $\mathfrak{L}$ has been found, one can find a unique set of structure constants ${C_{jk}}^i$ which captures the action of the Lie bracket. However, from equation (4.34) one can see that the structure constants of the Lie algebra are not necessarily constant, but instead may be smooth functions depending on the position in the manifold particular case. Thus, it would be interesting to study the Killing form which arises for this Lie algebra in detail.

4.3.2 Visualising Confidence Regions

Given that scaling with a non-vanishing smooth function does not affect the integral manifold associated with a set of vector fields, one can normalise their length without harm as $X \longmapsto \|X\|^{-1} X$. Since the local structural identifiability of a model guarantees that the likelihood-annihilating vector fields are non-vanishing (except of course at the MLE), one has that $\|X\| \neq 0$. Not only does this make it easier to visualise said vector fields directly, it can also significantly improve the stability of the numerical integration process. A visualisation of a likelihood-annihilating vector field is given in figure 33.

For high-dimensional parameter manifolds $\mathcal{M}$, one is generally limited to visualising two-dimensional

[59] Also, it should be noted that while it is necessary to assume that the second derivatives of the log-likelihood exist, it is not required that the second derivatives be non-vanishing (unlike the first derivatives) for this calculation to remain valid.

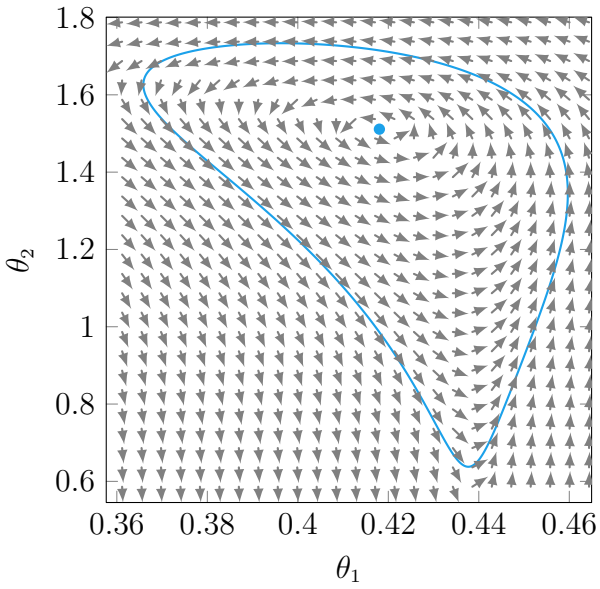


Figure 33: Visualisation of likelihood-annihilating vector field for non-linearly parametrised toy model specified in equation (5.2) as well as the 7σ confidence boundary which is obtained as its integral curve, given an appropriate initial condition in the form of a point which is already known to lie on the boundary of the 7σ confidence region.

or three-dimensional subspaces at a time. Typically, it is most straightforward to pick two-dimensional planes in $\mathcal{M}$ which one can specify, for example, via

$$P = \left\{ (\vec{p}_0 + s\vec{v}_1 + t\vec{v}_2) \in \mathbb{R}^{\dim \mathcal{M}} \;\middle|\; s, t \in \mathbb{R} \right\} \tag{4.43}$$

where the vectors $\vec{p}_0, \vec{v}_1, \vec{v}_2$ which define the plane are constant.[60] To visualise likelihood-annihilating vector fields in this slice, one can simply project them onto this plane while making sure that an element $\vec{\alpha} \in \mathcal{H}$ is chosen such that the orientation of the vector field remains consistent under the projection and that the projected vector field does not vanish in the plane. A simple example of an integral surface which is visualised via a family of integral curves is shown in figure 34.

4.3.3 Integral Manifolds and Frobenius' Theorem

Due to its broad applicability, several different formulations of Frobenius' theorem exist: e.g. in terms of vector fields, differential forms or so-called Pfaffian systems. However, each of them have the same statement at their core and can be shown to be equivalent. More in-depth discussions of Frobenius' theorem and its various formulations can be found, for example, in [42, 82]. Although Frobenius' theorem can be stated very succinctly, it uses some technical definitions which need to be introduced first:

A foliation $\mathcal{F}$ of a manifold $\mathcal{M}$ is a collection of mutually disjoint, connected, non-empty embedded

[60] To ensure that P indeed spans a plane, the basis vectors $\vec{v}_1$ and $\vec{v}_2$ must of course be linearly independent.

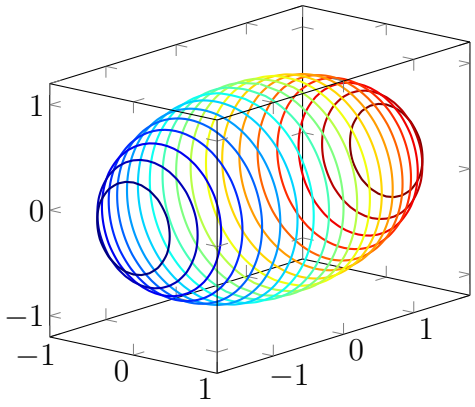


Figure 34: Schematic illustration of an ellipsoidal confidence region whose boundary is parametrised by a family of one-dimensional planar integral curves.

submanifolds of $\mathcal{M}$ whose union is all of $\mathcal{M}$, i.e. $\dot{\bigcup}_{U\in\mathcal{F}} U = \mathcal{M}$. The elements of $\mathcal{F}$ are also referred to as the leaves of the foliation.

A distribution $\mathfrak{D} \subseteq T\mathcal{M}$ on a smooth manifold $\mathcal{M}$ is a subbundle of the tangent bundle $T\mathcal{M}$, which means that it is described by specifying a linear subspace of each tangent space $\mathfrak{D}_p \subseteq T_p\mathcal{M}$ for all points $p \in \mathcal{M}$. Further, a distribution $\mathfrak{D}$ is said to be involutive if it is closed with respect to the Lie bracket, i.e. if $X, Y \in \mathfrak{D} \implies [X, Y] \in \mathfrak{D}$. In other words, the vector fields specified by $\mathfrak{D}$ must form a Lie subalgebra of $\big(\Gamma(T\mathcal{M}), [\,\cdot\,,\,\cdot\,]\big)$.[61]

A (non-empty) submanifold $\mathcal{N} \subseteq \mathcal{M}$ is the integral manifold of a distribution $\mathfrak{D}$ if at each point $p \in \mathcal{N}$ one has $\mathfrak{D}_p = T_p\mathcal{N}$, which means that the distribution $\mathfrak{D}$ is precisely the tangent bundle of the submanifold $\mathcal{N}$. In particular, it being a submanifold implies that $\mathcal{N}$ must have the same dimension at every point $p \in \mathcal{N}$, wherefore self-intersections of submanifolds $\mathcal{N} \subseteq \mathcal{M}$ are categorically excluded (see also figure 14). Intuitively, one can think of an n-dimensional integral manifold associated with a set of vector fields as being meshed by the family of integral curves to said vector fields.

While it is clear that any smooth submanifold $\mathcal{N} \subseteq \mathcal{M}$ has a unique distribution $\mathfrak{D}$ as its tangent bundle $T\mathcal{N}$, the converse is not obvious from the outset. However, this is precisely the assertion made by Frobenius' theorem which states:

A distribution $\mathfrak{D} \subseteq T\mathcal{M}$ is integrable (i.e. generates a unique family of integral manifolds) if and only if $\mathfrak{D}$ forms a closed Lie subalgebra of $\big(\Gamma(T\mathcal{M}), [\,\cdot\,,\,\cdot\,]\big)$, i.e. if $\mathfrak{D}$ is involutive.

It is easy to see that in the case of one-dimensional integral curves, which are generated by (the span of) a single vector field, the condition that it should commute with itself is trivially fulfilled. Therefore, Frobenius' theorem coincides with the Picard–Lindelöf theorem on the existence and uniqueness of integral curves and can be thought of as a higher-dimensional generalisation of the Picard–Lindelöf theorem in the sense that it provides a necessary and sufficient condition for the existence of unique integral surfaces and manifolds.

[61] While it is possible to make this slightly more general, the discussion here specifically focuses on smooth distributions (i.e. smooth subbundles of $T\mathcal{M}$).

Furthermore, since the family of integral manifolds generated by a distribution is unique, they must be mutually non-intersecting. Therefore, assuming that the distribution is defined on all of $\mathcal{M}$, its integral manifolds foliate $\mathcal{M}$.

4.4 Wilks' theorem

Since the likelihood depends on data points which are subject to statistical errors and noise, the likelihood itself also constitutes a random variable in the frequentist view. As such, it is assumed that the likelihood must therefore also have some (possibly unknown) underlying probability distribution.

A well-established theorem on the distribution of the likelihood ratio is due to S. Wilks (see [84]) and can be summarised as follows:

In the limit that one has a large number of data points, the difference in log-likelihoods is asymptotically distributed according to

$$-2\Big(\ell(\theta) - \ell(\theta_{\text{MLE}})\Big) \sim \chi^2_k \qquad (\text{as } N \longrightarrow \infty) \tag{4.44}$$

with deviations on the order of $O(1/\sqrt{N})$ and where k denotes the number of components in which θ and θ_{MLE} differ, i.e. the degrees of freedom between the hypotheses corresponding to the parameters θ_{MLE} and θ.

Importantly, the proof of this statement assumes that θ_{MLE} is the "true" parameter configuration underlying the observations. The popularity of the likelihood ratio test and by extension the significance of this theorem is due to the Neyman–Pearson lemma (see [51]), which asserts that the likelihood ratio (and thus also the difference in log-likelihoods) is at least as powerful in discriminating between so-called simple hypotheses as any other test statistic.

In cases where the number of data points N is too small for Wilks' theorem to yield a good approximation, it may be advisable to use an alternative such as the F-test to discriminate between hypotheses (see section 4.1.2) instead of the likelihood ratio test.[62]

4.5 Pointwise Confidence Bands

In most publications (see e.g. [43]), one finds a definition of confidence bands along the following lines: Two functions $l(x)$ and $u(x)$ constitute the boundary of a pointwise confidence band associated with the confidence level q around a model $y_{\text{model}}(x;\theta_{\text{MLE}})$ if

$$\forall x \in \mathcal{X}: \qquad \mathbb{P}\Big(l(x) \leq y_{\text{model}}(x;\theta_{\text{MLE}}) \leq u(x)\Big) = q. \tag{4.45}$$

That is, at each $x \in \mathcal{X}$, the interval $[l(x), u(x)] \subseteq \mathcal{Y}$ separately provides a confidence interval around the prediction $y_{\text{model}}(x;\theta_{\text{MLE}})$ of the model function. Importantly, pointwise confidence

[62] Rigorous upper bounds on the asymptotic normality of likelihoods are discussed in [6, 7].

bands are not to be confused with simultaneous confidence bands which, in contrast, are defined as

$$\mathbb{P}\Big(\forall x \in \mathcal{X} : \quad l(x) \leq y_{\text{model}}(x; \theta_{\text{MLE}}) \leq u(x)\Big) = q \tag{4.46}$$

which differs only subtly from the definition of pointwise confidence bands in its placement of the "$\forall x \in \mathcal{X}$" qualification.

Apart from the fact that the definition of pointwise confidence bands in equation (4.45) is only applicable for one-dimensional observations, i.e. when $\dim \mathcal{Y} = 1$, it also does not provide a practical recipe with which to calculate said confidence bands. Thus, a better definition of a pointwise confidence band of level q is arguably given by

$$\mathcal{B}_q(x) \coloneqq y_{\text{model}}(x; \mathcal{C}_q) = \Big\{ y_{\text{model}}(x; \theta) \in \mathcal{Y} \;\Big|\; \theta \in \mathcal{C}_q \Big\} \tag{4.47}$$

which generalises to higher dimensional observation spaces (i.e. $\dim \mathcal{Y} > 1$) much more gracefully. Here, $\mathcal{B}_q(x) \subseteq \mathcal{Y}$ specifies a set of predictions which is estimated to contain the mean of observations which are made at the conditions x with a probability of q.

A definition of pointwise confidence bands in this manner also has the benefit of not presupposing any particular form for the error distribution (e.g. a normal distribution) around the model. Instead, any general error distribution is already incorporated into the confidence regions $\mathcal{C}_q$ via the likelihood function $L(\text{data} \,|\, \theta)$. Therefore, the confidence bands $\mathcal{B}_q$ are not affected by non-linear distortions of confidence regions due to the model parametrisation. This also means that the interpretation of pointwise confidence bands $\mathcal{B}_q$ has a similar frequentist interpretation to confidence regions when defined in this manner: It estimates that if the experiment were to be repeated many times with different datasets, the probability for the model prediction associated with the maximum likelihood to lie within the confidence band $\mathcal{B}_q$ at any given $x \in \mathcal{X}$ is q.

Since the confidence regions $\mathcal{C}_q$ are established by virtue of the likelihood ratio, whose value is independent of the parametrisation of the model y_{model}, the confidence bands $\mathcal{B}_q$ are also independent of the model parametrisation. In this sense, the best possible model prediction $y_{\text{model}}(x; \theta_{\text{MLE}})$ and the confidence bands are both examples of "objective", i.e. invariant quantities.

Just as with confidence regions, one is often particularly interested in finding the boundary of a confidence band for the purpose of illustration. That is, one wishes to draw curves or surfaces, which bound all possible predictions of a model below a confidence level q, meaning that they encompass the prediction of the true model underlying the data with a confidence level q.

To parametrise the boundary of the pointwise confidence band $(\partial \mathcal{B}_q)(x) = \partial\Big(y_{\text{model}}(x; \mathcal{C}_q)\Big)$, it may suffice for some models to calculate $y_{\text{model}}(x; \partial \mathcal{C}_q)$ which is computationally less expensive, since fewer evaluations of the model are necessary. While it intuitively makes sense that the most extreme deviations from the maximum likelihood prediction should be attained for parameter configurations on the boundary of the confidence region $\mathcal{C}_q$ and explicit numerical tests of realistic models appear to confirm this, proving this rigorously for generic models is not entirely straightforward.

Specifically, for a map $y : \mathcal{M} \longrightarrow \mathcal{Z}$ and some set $C \subseteq \mathcal{M}$, one would like to prove the

topological relation

$$\partial\big(y(C)\big) \subseteq y(\partial C) \tag{4.48}$$

under the weakest assumptions possible. It is reasonable to assume that the map y is continuously differentiable with respect to the parameter $\theta \in \mathcal{M}$ and that C is a compact set. As elaborated on in section 3.2, it is only assumed that the model is continuous with respect to the conditions $x \in \mathcal{X}$, which means the case

$$\mathcal{Z} = C^0(\mathcal{X}, \mathcal{Y}) = \{f : \mathcal{X} \longrightarrow \mathcal{Y} \mid f \text{ continuous}\} \tag{4.49}$$

is considered here.

A detailed proof is given in the appendix, which shows that sufficient conditions for equation (4.48) to hold are that the map $y : \mathcal{M} \longrightarrow C^0(\mathcal{X}, \mathcal{Y})$ be injective as well as continuous and that the set $C \subseteq \mathcal{M}$ be compact. The injectivity and continuity of the model map y_{model} are satisfied if it is globally structurally identifiable. Moreover, since $C \subseteq \mathcal{M}$ represents the confidence region $\mathcal{C}_q$ which is typically bounded for $q < 1$ if y_{model} is globally structurally identifiable, its compactness is also given.

When applicable, equation (4.48) represents a significant reduction in computational effort not only because fewer evaluations of the likelihood are necessary but also since Monte Carlo methods which sample the interior of the confidence region and the associated loss of accuracy can consequently be avoided. Most importantly, the benefit of defining pointwise confidence bands through equation (4.47) is that they are obtained almost "for free" once the confidence region $\mathcal{C}_q$ is known. In particular, the integral manifold method described in section 4.3 provides a convenient parametrisation of confidence boundaries $\partial\mathcal{C}_q$ such that the model can be efficiently evaluated for all parameters $\theta \in \partial\mathcal{C}_q$ at any desired $x \in \mathcal{X}$ to establish the simultaneous confidence bands.

This is also the reason why this section focuses on pointwise confidence bands instead of simultaneous confidence bands: in contrast to pointwise confidence bands, it is not precisely clear how simultaneous confidence bands are connected to the confidence regions $\mathcal{C}_q$ and therefore simultaneous confidence bands generally require a more elaborate calculation.

Since the confidence bands are spread apart further for x-values where the uncertainty about the model prediction is larger, one can conversely use them to qualitatively judge under which conditions new observations will contribute the most useful information. One might for example wish to visualise the width of a confidence band (for some level q) as a function of the observation conditions via

$$w(x) = u(x) - l(x). \tag{4.50}$$

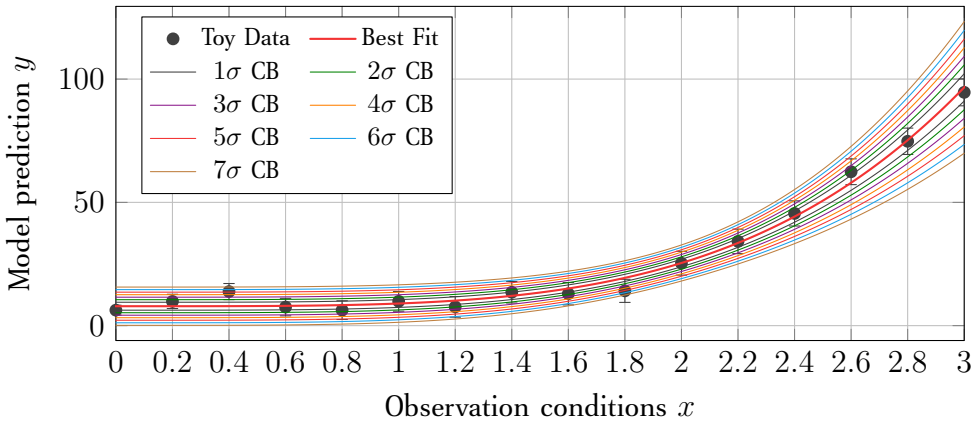


Figure 35: This plot illustrates the pointwise confidence bands ranging from 1σ to 7σ around the maximum likelihood model prediction of the toy example specified in equation (5.1). As is characteristic for any confidence band, one can see that the confidence bands more narrowly enclose the maximum likelihood prediction for values $x \in \mathcal{X}$ that lie in the "midrange" of the dataset.

4.6 Rewriting the Log-Likelihood Difference

For a one-dimensional parameter space $\mathcal{M}$, the difference between two log-likelihoods can be rewritten by virtue of the fundamental theorem of calculus to yield

$$\ell(\theta_1) - \ell(\theta_2) = \int_{\partial[\theta_1,\theta_2]} \ell = \int_{[\theta_1,\theta_2]} \mathrm{d}\ell = \int_{[\theta_1,\theta_2]} \mathrm{d}\tilde{\theta}\, \ell'(\tilde{\theta}) = \int_{[0,1]} \mathrm{d}t\, |\theta_2 - \theta_1|\, \ell'\Big((\theta_2 - \theta_1)t + \theta_1\Big) \tag{4.51}$$

using the substitution $\tilde{\theta}(t) = (\theta_2 - \theta_1)t + \theta_1$ in the last equality. As a result, the log-likelihood difference is computed by summing up the score along the way between the parameter values.

To extend this idea to parameter manifolds of dimension two or higher, the fundamental theorem of calculus is replaced by Stokes' theorem which assumes that $\mathcal{M}$ is orientable and that the boundary of the domain of integration is piecewise smooth. Additionally, since there are now a multitude of paths connecting any two given parameter configurations one must specify a (smooth) curve $\gamma : I \longrightarrow \gamma(I) \subseteq \mathcal{M}$ with $I \subseteq \mathbb{R}$ connected[63] and $\gamma(\partial I) = \{\theta_1, \theta_2\}$, i.e. some smooth curve connecting the points $\theta_1, \theta_2 \in \mathcal{M}$. Given this, one can rewrite the difference as

$$\ell(\theta_1) - \ell(\theta_2) = \int_{\partial(\gamma(I))} \ell = \int_{\gamma(I)} \mathrm{d}\ell = \int_I \gamma^*(\mathrm{d}\ell). \tag{4.52}$$

The integrand can now be evaluated as

$$\gamma^*(\mathrm{d}\ell) = \mathrm{d}(\gamma^*\ell) = \mathrm{d}(\ell \circ \gamma) \tag{4.53}$$

[63] The connectedness of I is necessary to ensure that $\gamma(I) \subseteq \mathcal{M}$ is connected and as a result only has two boundary points.

or equivalently by expressing the score $\mathrm{d}\ell$ in a chart (U, θ) as $\mathrm{d}\ell = \omega_a \, \mathrm{d}\theta^a$

$$\gamma^*(\omega_a \, \mathrm{d}\theta^a) = \gamma^*(\omega_a) \wedge \gamma^*(\mathrm{d}\theta^a) = (\omega_a \circ \gamma) \; \mathrm{d}(\gamma^* \theta^a) = (\omega_a \circ \gamma) \; \mathrm{d}(\theta^a \circ \gamma) = (\omega_a \circ \gamma) \, \dot{\gamma}^a \tag{4.54}$$

and thus one finally arrives at

$$\ell(\theta_1) - \ell(\theta_2) = \int_I \mathrm{d}t \, \dot{\gamma}^a(t) \, (\omega_a \circ \gamma)(t) = \int_I \mathrm{d}t \, \dot{\gamma}^a(t) \, \frac{\partial \ell}{\partial \theta^a}\bigg|_{\gamma(t)} \tag{4.55}$$

which constitutes the higher-dimensional analogue to equation (4.51).

4.7 Relation of Geodesic Length to Confidence Intervals

By squaring the integral representation of the log-likelihood difference from equation (4.55) and subsequently taking the expectation with respect to the observation y_{data} on both sides, one can calculate

$$\mathbb{E}_{y_{\mathrm{data}}}\Big(\big(\ell(\theta_1) - \ell(\theta_2)\big)^2\Big) \equiv \mathbb{E}_{y_{\mathrm{data}}}\Big(\big(\ell(\mathrm{data} \,|\, \theta_1) - \ell(\mathrm{data} \,|\, \theta_2)\big)^2\Big) \tag{4.56}$$

$$= \mathbb{E}_{y_{\mathrm{data}}}\left(\left[\int_I \mathrm{d}t \, 1 \cdot \dot{\gamma}^a(t) \, \frac{\partial \ell}{\partial \theta^a}\big(\gamma(t)\big)\right]^2\right) \leq \mathbb{E}_{y_{\mathrm{data}}}\left(\left[\int_I \mathrm{d}t \, 1^2\right]\left[\int_I \mathrm{d}t \left(\dot{\gamma}^a(t) \, \frac{\partial \ell}{\partial \theta^a}\big(\gamma(t)\big)\right)^2\right]\right) \tag{4.57}$$

$$= \mathrm{vol}(I) \, \mathbb{E}_{y_{\mathrm{data}}}\left(\int_I \mathrm{d}t \left(\dot{\gamma}^a(t) \, \frac{\partial \ell}{\partial \theta^a}\big(\gamma(t)\big)\right)^2\right) = \mathrm{vol}(I) \int_I \mathrm{d}t \; \mathbb{E}_{y_{\mathrm{data}}}\left(\left(\dot{\gamma}^a(t) \, \frac{\partial \ell}{\partial \theta^a}\big(\gamma(t)\big)\right)^2\right) \tag{4.58}$$

$$= \mathrm{vol}(I) \int_I \mathrm{d}t \; \mathbb{E}_{y_{\mathrm{data}}}\left(\dot{\gamma}^a(t) \, \frac{\partial \ell}{\partial \theta^a}\big(\gamma(t)\big) \, \frac{\partial \ell}{\partial \theta^b}\big(\gamma(t)\big) \dot{\gamma}^b(t)\right) \tag{4.59}$$

$$= \mathrm{vol}(I) \int_I \mathrm{d}t \; \underbrace{\mathbb{E}_{y_{\mathrm{data}}}\left(\frac{\partial \ell}{\partial \theta^a}\big(\gamma(t)\big) \, \frac{\partial \ell}{\partial \theta^b}\big(\gamma(t)\big)\right)}_{\equiv g_{ab}\left(\gamma(t)\right)} \dot{\gamma}^a(t) \, \dot{\gamma}^b(t) = \mathrm{vol}(I) \int_I \mathrm{d}t \, g_{\gamma(t)}\big(\dot{\gamma}(t), \dot{\gamma}(t)\big) \tag{4.60}$$

$$= \mathrm{vol}(I) \, E[\gamma] \tag{4.61}$$

where the Cauchy–Schwarz inequality was used in the second line.[64] Crucially, this derivation assumes that the order of integration between the integral over the curve parameter t and the integral involved in taking the expectation over y_{data} can be interchanged.

Also employing equation (2.87), the overall result can be summarised as

$$\mathbb{E}_{y_{\mathrm{data}}}\Big(\big(\ell(\theta_1) - \ell(\theta_2)\big)^2\Big) \overset{(1)}{\leq} \big(L[\gamma]\big)^2 \overset{(2)}{\leq} \mathrm{vol}(I) \, E[\gamma]. \tag{4.62}$$

In the case that γ is an affinely parametrised geodesic (i.e. a minimiser of E), the second inequality (2) in this statement becomes an equality, i.e. $L[\gamma]^2 = \mathrm{vol}(I) \, E[\gamma]$. More importantly, the

[64] Specifically, the integral over the curve parameter t is interpreted as an inner product on the space of continuous functions over the curve domain $C^0(I)$.

first inequality (1) becomes an equality whenever the constant one-function and the function $\dot{\gamma}^a(t)\,\partial\ell/\partial\theta^a|_{\gamma(t)}$ are linearly dependent. Formally, one must have

$$\exists \alpha, \beta \in \mathbb{R} \setminus \{0\} : \forall t \in I : \qquad \alpha \cdot 1 = \beta \cdot \dot{\gamma}^a(t)\, \frac{\partial \ell}{\partial \theta^a}\Big(\gamma(t)\Big). \tag{4.63}$$

In other words, the combination $\dot{\gamma}^a(t)\,\partial\ell/\partial\theta^a|_{\gamma(t)}$ must be constant for all values of t. While the direction of affinely parametrised geodesics may change in coordinates, the length of the tangent vector $\|\dot{\gamma}\|$ remains constant. To emphasise that γ is assumed to be a geodesic connecting θ_1 and θ_2 from this point on, the geodesic length of γ will be denoted by $d_{\mathcal{M}}(\theta_1, \theta_2) = L[\gamma]$ where the metric function $d_{\mathcal{M}}$ is the same as in equation (2.78).

Since $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$, one can rewrite the expectation of the squared log-likelihood difference in equation (4.62) as

$$\mathbb{E}\Big((\ell_1 - \ell_2)^2\Big) = \mathbb{E}(\ell_1 - \ell_2)^2 + \mathrm{Var}(\ell_1 - \ell_2) \tag{4.64}$$

from which it follows that

$$\mathbb{E}_{y_{\text{data}}}\Big(\ell(\theta_1) - \ell(\theta_2)\Big)^2 + \mathrm{Var}_{y_{\text{data}}}\Big(\ell(\theta_1) - \ell(\theta_2)\Big) \leq d_{\mathcal{M}}(\theta_1, \theta_2)^2 \tag{4.65}$$

which constitutes the final result.

Although I am unable to find a mistake in the logic of this derivation, it appears that one can find numerical counter-examples where this inequality does not hold. At this point, it is not known whether these counter-examples are a result of the non-interchangeability of the order of integration or whether they violate some other assumption made in the derivation. Therefore, more detailed investigations into the veracity of this inequality are necessary.

If true, this result would allow for the incorporation of knowledge about the distribution of the log-likelihood ratio test in comparisons with geodesic distance. Using Wilks' theorem, i.e. under the assumption that one of the two hypotheses is true, say $\theta_2 = \theta_{\text{MLE}}$, one finds in the large sample limit

$$\mathrm{Var}\Big(-2(\ell_1 - \ell_2)\Big) = 4\mathrm{Var}\Big((\ell_1 - \ell_2)\Big) \sim \mathrm{Var}\Big(\chi^2_k\Big) \qquad (\text{as } N \longrightarrow \infty) \tag{4.66}$$

in a slight abuse of notation on the right-hand side. This step is permissible since it can be shown that asymptotic relations remain valid under integration, provided that the integral of both sides exists (see [13]). Therefore, the fact that the log-likelihood difference is asymptotically distributed according to a χ^2-distribution also implies that the variance of the log-likelihood difference is asymptotic to the variance of the χ^2-distribution as $N \longrightarrow \infty$. This immediately implies

$$\mathrm{Var}\Big((\ell_1 - \ell_2)\Big) \sim \frac{1}{4}\mathrm{Var}\Big(\chi^2_k\Big) = \frac{2k}{4} = \frac{k}{2} \qquad (\text{as } N \longrightarrow \infty). \tag{4.67}$$

Therefore, assuming a large dataset, the general result from equation (4.65) can be specialised to yield

$$\mathbb{E}\Big(\ell(\theta_1) - \ell(\theta_{\text{MLE}})\Big)^2 + k/2 \leq d_{\mathcal{M}}(\theta_1, \theta_{\text{MLE}})^2. \tag{4.68}$$

Unfortunately, a non-trivial lower bound for the expected log-likelihood difference has remained elusive thus far. If there are no restrictions on the form of ℓ, it is intuitively clear that such a lower bound should not exist. For example, given a model parametrisation for which the parameter manifold $\mathcal{M}$ features a mirror symmetry, then the likelihood may have two peaks arbitrarily far apart. The geodesic distance between the two peaks would then be large while the log-likelihood difference vanishes. On the other hand, if the log-likelihood ℓ is known to be globally structurally identifiable or concave, and one of the parameter configurations corresponds to the MLE, it is at least plausible that such a bound might exist.

4.8 Geodesic Distance as a Hypothesis Test

A further hypothesis test not discussed in section 4.1 is the so-called Wald test which is given by

$$W \coloneqq (\theta - \theta_{\text{MLE}})^a \, g_{ab} \, (\theta - \theta_{\text{MLE}})^b \sim \chi^2_k \qquad (\text{as } N \longrightarrow \infty) \tag{4.69}$$

where g denotes the Fisher information and $\theta \in \mathcal{M}$ is any parameter configuration. As with most other hypothesis tests, it is assumed that one of the hypotheses is already known to be "true", which is denoted via θ_{MLE} here. Lastly, k is the number of degrees of freedom, i.e. the number of components in which the two parameter configurations differ.

Due to the fact that it is derived by expanding the likelihood ratio test in a Taylor series, the Wald test is no longer invariant under reparametrisation which may be seen as its biggest disadvantage. Furthermore, compared with the likelihood ratio test, whose only assumption is that the number of data points is large enough for the χ^2_k-distribution to be a yield approximation of the distribution of the likelihood ratio (i.e. Wilks' theorem), the Wald test additionally assumes that the covariance between the parameters is given by the inverse Fisher metric. In other words, it assumes that the Cramér–Rao lower bound is attained.

However, it is not difficult to recognise that the Wald statistic corresponds precisely to the square of the geodesic distance between the two parameters $\theta, \theta_{\text{MLE}} \in \mathcal{M}$ in the case that the Fisher metric g is constant. In other words,

$$W = d_{\mathcal{M}}(\theta_{\text{MLE}}, \theta)^2. \tag{4.70}$$

Given that this relationship is valid in coordinates where the Fisher metric is constant and geodesic length is invariant under reparametrisation, it therefore follows that while equation (4.69) no longer provides an accurate measure of distance of a parameter configuration to the MLE for non-linearly parametrised models, the geodesic distance does. One might therefore conjecture that this relationship is also valid in coordinate systems where the Fisher metric is not necessarily constant but the parameter manifold $\mathcal{M}$ is still flat.

Indeed, extensive numerical tests appear to confirm that for normal likelihoods the geodesic distance of points on the boundary of a confidence region to the MLE is constant even for non-linearly parametrised models. An example of this can be found in figure 43 and table 2. Therefore, the squared geodesic distance can also be used as an approximation to the likelihood ratio test. As a result, it is possible to define confidence regions in a completely geometric way (see section 4.9).

Although the results obtained by using the Wald statistic may be less accurate than the the likelihood ratio test which it approximates, it has the potential of significantly reducing the computational effort involved in the computation of confidence regions. That is because the data space $\mathcal{D}$ may be very high-dimensional compared with $\mathcal{M}$.

While it is already possible to read off from equation (4.69) how the geodesic distance should be related to the the confidence level, one may alternatively derive this as follows:

In spherical coordinates, the volume element on $\mathbb{R}^N$ can be expressed as

$$\mathrm{d}^N x = r^{N-1}\,\mathrm{d}r\,\sin^{N-2}(\vartheta_1)\,\mathrm{d}\vartheta_1\,\sin^{N-3}(\vartheta_2)\,\mathrm{d}\vartheta_2\,...\,\sin^{1}(\vartheta_{N-2})\,\mathrm{d}\vartheta_{N-2}\,\mathrm{d}\varphi = r^{N-1}\,\mathrm{d}r\,\mathrm{d}\Omega \tag{4.71}$$

where $r \in \mathbb{R}_0^+$, $\vartheta_1, ..., \vartheta_{N-2} \in [0, \pi]$ and $\varphi \in [0, 2\pi)$. By requiring that the volume contained under a multivariate normal distribution of a covariance ellipse with radius R should be given by some $q \in [0, 1]$, one can work out that

$$1 - q \overset{!}{=} \int\limits_{\mathbb{R}^N \setminus B_R(0)} \mathrm{d}^N x\, p(x; 0, \mathbb{1}_N) = \Big((2\pi)^N\,\det(\mathbb{1}_N)\Big)^{-\frac{1}{2}} \int\limits_{R<\|x\|} \mathrm{d}^N x \exp\left(-\frac{1}{2}\sum_{j=1}^{N}(x_j)^2\right) \tag{4.72}$$

$$= (2\pi)^{-\frac{N}{2}} \underbrace{\int\limits_{S^{N-1}} \mathrm{d}\Omega}_{2\pi^{N/2}/\Gamma(N/2)} \underbrace{\int\limits_{R}^{\infty} \mathrm{d}r\, r^{N-1} \exp\left(-\frac{r^2}{2}\right)}_{\equiv 2^{N/2}\Gamma(N/2,\, R^2/2)} = \frac{\Gamma\left(\frac{N}{2}, \frac{R^2}{2}\right)}{\Gamma\left(\frac{N}{2}\right)} \equiv Q\left(\frac{N}{2}, \frac{R^2}{2}\right) \tag{4.73}$$

where $Q(a, b)$ is the so-called "regularised upper incomplete gamma function". Thus, using

$$1 - q = 1 - \mathrm{erf}\left(\frac{n}{\sqrt{2}}\right) = \mathrm{erfc}\left(\frac{n}{\sqrt{2}}\right) \tag{4.74}$$

one finds that the relationship between the geodesic radius R and the confidence level n for an N-dimensional normal distribution is given by

$$d(\theta_{\text{MLE}}, \theta_{n\sigma}) = R = \sqrt{2\,Q^{-1}\left(\frac{N}{2}, \mathrm{erfc}\left(\frac{n}{\sqrt{2}}\right)\right)} \tag{4.75}$$

It was already observed for example by B. Schäfer in [63] that in the case of $\dim \mathcal{M} = N = 2$, one finds the analytic expression

$$d_{\mathcal{M}}(\theta_{\text{MLE}}, \theta)^2 = -2 \ln \mathrm{erfc}\left(\frac{n}{\sqrt{2}}\right) \tag{4.76}$$

which is consistent with equation (4.75) since one can show that $Q(1, r^2/2) = \exp\big(-r^2/2\big)$. Incidentally, one can verify that the cumulative distribution function of the χ_k^2-distribution is precisely given by the regularised lower incomplete gamma function $P(a, b)$ such that they satisfy

together $P(a,b) + Q(a,b) = 1$ for all $a \in \mathbb{R}^+$ and $b \in \mathbb{R}^+$. Therefore, it is possible to rewrite equation (4.75) as

$$d_{\mathcal{M}}(\theta_{\text{MLE}}, \theta_{n\sigma}) = \sqrt{F_N^{-1}\left(\operatorname{erf}\left(\frac{n}{\sqrt{2}}\right)\right)} \tag{4.77}$$

where in this case F_N^{-1} is the quantile function for a χ_k^2-distribution with N degrees of freedom. Finally, since Wilks' theorem states $2\Big(\ell(\theta_{\text{MLE}}) - \ell(\theta)\Big) \sim \chi_k^2$ and it is also true that $d_{\mathcal{M}}(\theta_{\text{MLE}}, \theta_{n\sigma}) \sim \chi_k^2$, it is not hard to see that one effectively has the asymptotic relation

$$2\Big(\ell(\theta_{\text{MLE}}) - \ell(\theta)\Big) \sim d_{\mathcal{M}}(\theta_{\text{MLE}}, \theta_{n\sigma}) \qquad (\text{as } N \longrightarrow \infty) \tag{4.78}$$

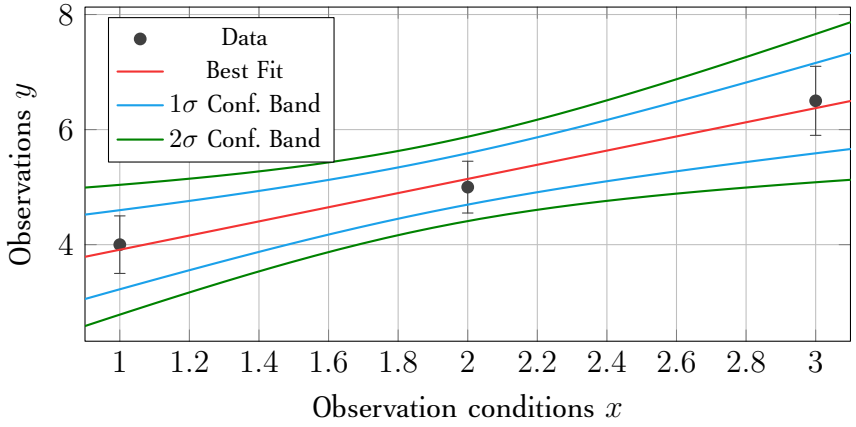
in the case of normal likelihoods. Therefore, the geodesic distance from the MLE may be used as an approximation of the log-likelihood difference on the parameter manifold in the large sample limit $N \longrightarrow \infty$. Figure 43 and table 2 demonstrate that the geodesic distance from the MLE to likelihood-based confidence boundaries of any confidence level can be accurately predicted using equation (4.75). This is further evidence in favour of the conjecture that the squared geodesic distance yields the correct generalisation to the Wald test.

4.9 Proposed Algorithm for Construction of Confidence Regions

In section 4.8, it is discussed that the geodesic distance yields a very accurate approximation to the likelihood ratio test for normal likelihoods. As a result, the following algorithm is proposed for the construction of confidence regions for datasets with normal error distribution:

1. Compute the geodesic distance of the MLE to the confidence boundary of interest via equation (4.75).
2. Construct a geodesic emanating from the MLE in any direction up to this length.
3. Use the flows generated by likelihood-annihilating vector fields to parametrise the level set associated with the end point of the geodesic.

The advantage of searching for a point on the boundary of the confidence region along a geodesic compared with evaluating a hypothesis test such as the log-likelihood ratio along a coordinate axis is that the geodesic naturally accounts for non-linear distortions of the coordinates introduced by the parametrisation of the model. Overall, this generally leads to a significant reduction in the number of evaluations of the likelihood that are necessary to locate the confidence boundary with sufficient precision. Since the likelihood can be very expensive to evaluate for large datasets, this should increase the performance of the confidence boundary construction. Moreover, this results in a completely geometric and parametrisation-invariant definition of confidence regions and by extension also confidence bands. If very high accuracy is necessary, the estimate for the location of the confidence boundary obtained from the geodesic distance may of course be further refined using the likelihood ratio test.



x	y	σ
1	4	0.5
2	5	0.45
3	6.5	0.6

Figure 36: A brief summary of the dataset on which the confidence regions from figure 39 are based is provided. In addition to the data and best fit, the left-hand side plot depicts the pointwise confidence bands of level 1σ and 2σ generated from the confidence boundaries. The right-hand side table details the exact values of the artificial dataset.

4.10 Qualitative Effects of Reparametrisation on Confidence Regions

The aim of this section is to provide a small survey of the qualitative effects that model reparametrisations can have on the shape of confidence regions. A discussion about the suitability of the employed parametrisations such as their invertibility, differentiability, valid chart domains and so on is foregone in these examples.

The dataset which is used for this purpose consists of only three points and is illustrated in figure 36. Each of the various model parametrisations shown in figure 39 corresponds to the choice of a different chart on the same embedded prediction surface $h(U) \subseteq \mathcal{D}$ which encodes a linear relationship between x and y. As a result, while the 1σ and 2σ confidence regions for these parametrisations exhibit different distortions in coordinates, their image under the corresponding embedding map $h(\mathcal{C}_{1\sigma})$ is the same.

Figures 37 and 38 illustrate the two-dimensional surface corresponding to the parameter manifold embedded into the three-dimensional data space $\mathcal{D}$ under the map h. In both cases, the parameter space was sampled on a uniform grid to construct the embedded surface. On the one hand, figure 37 illustrates that the linearly parametrised embedding map preserves the linear spacing between evaluations on the parameter manifold. On the other hand, figure 37 shows that the initially uniform grid is distorted under the non-linearly parametrised embedding h. However, both surfaces constitute subsets of the same plane $h(\mathcal{M}) \subseteq \mathcal{D}$ which is merely coordinatised differently. The reason for this is that in both instances of the model, the resulting prediction constitutes a straight line (i.e. a linear relationship between x and y) irrespective of the chosen chart on $\mathcal{M}$. Since y_{data} does not lie on the embedded surface in either case, there is no possible parameter configuration for which the model predictions will perfectly reproduce the observations. Of course, to visualise data spaces with $\dim \mathcal{D} > 3$, some dimensions in the data space must be suppressed.

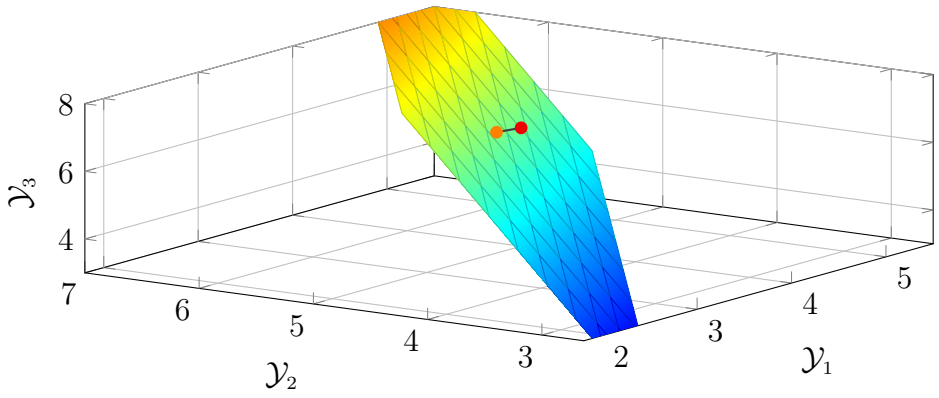


Figure 37: Visualisation of the three-dimensional data space $\mathcal{D}$ associated with the data from figure 36 including a part of the embedded parameter manifold $h(U) \subseteq \mathcal{D}$ corresponding to the linearly parametrised model $y_{\text{model}}(x; a, b) = a\,x + b$. The red point indicates the collective observations y_{data} whereas the orange point corresponds to the collective prediction of the model evaluated at the best fit, i.e. $h(\theta_{\text{MLE}}) \in \mathcal{D}$.

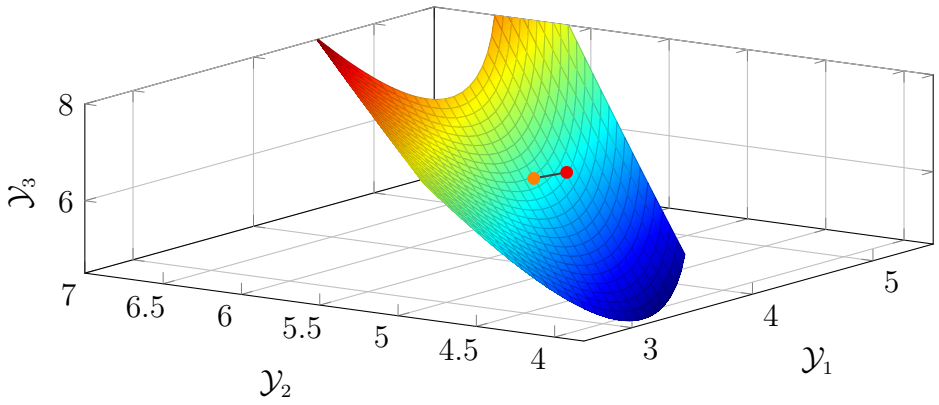


Figure 38: Visualisation of the three-dimensional data space $\mathcal{D}$ associated with the data from figure 36 including a part of the embedded parameter manifold $h(U) \subseteq \mathcal{D}$ corresponding to the model $y_{\text{model}}(x; a, b) = (a + b)\,x + \exp(a - b)$. The red point indicates the collective observations y_{data} whereas the orange point corresponds to the collective prediction of the model evaluated at the best fit, i.e. $h(\theta_{\text{MLE}}) \in \mathcal{D}$.

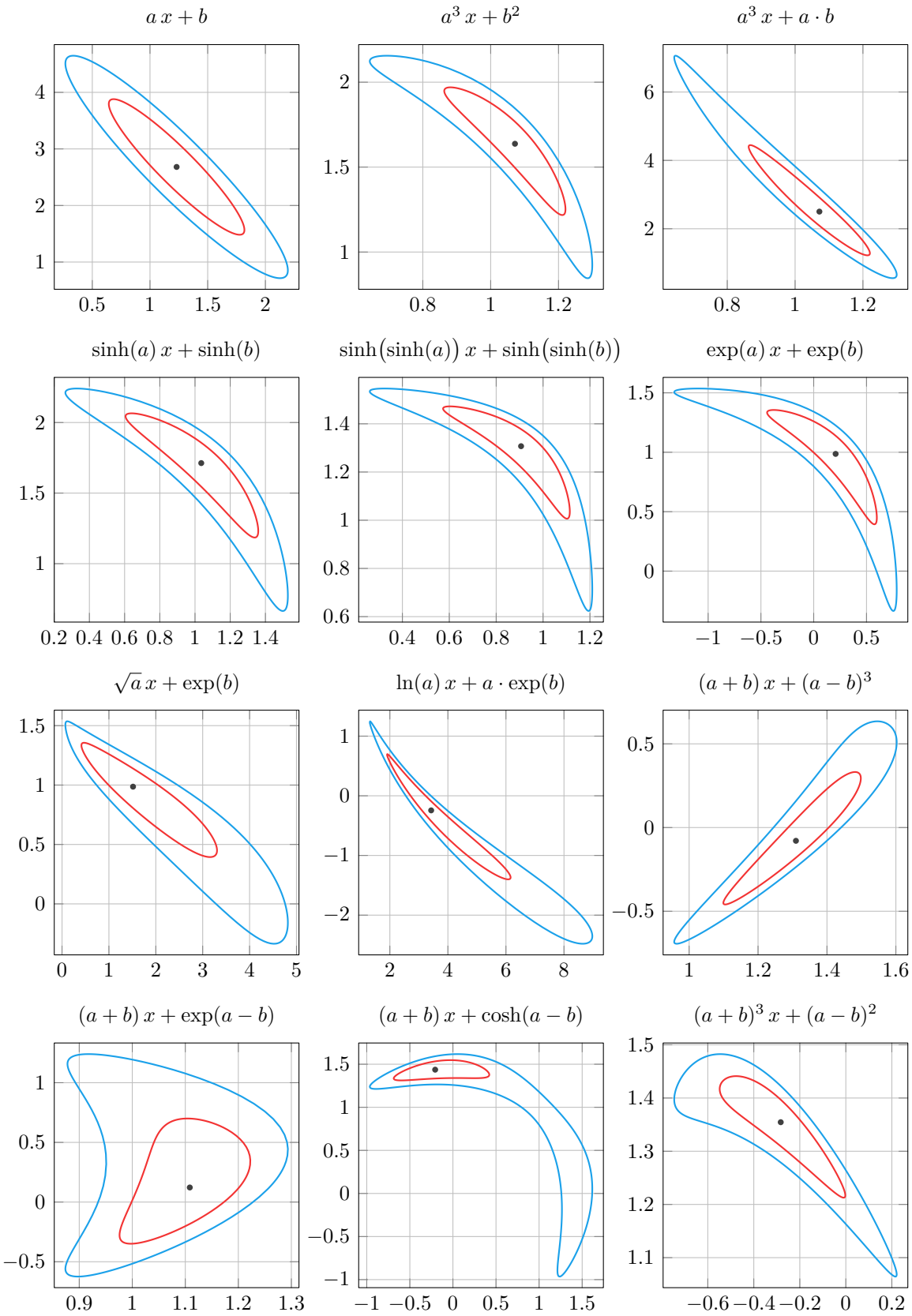


Figure 39: Illustration of confidence regions of level 1σ and 2σ for various alternative parametrisations (shown above the respective plots) of a model $y(x; a, b)$. For each of the chosen parametrisations, the model prediction consists of a straight line which is fitted to the dataset shown in figure 36.

5 Applications of Information Geometry

This section investigates some of the practical challenges faced in applications of the previously discussed information-geometric methods such as the computation of exact confidence regions. Furthermore, the performance, accuracy and numerical robustness of the developed schemes are compared against conventional methods.

5.1 Toy Model

Before applying the new methods developed in this thesis to cosmological data and real-world problems, it is advisable to first demonstrate the concepts on structurally simple toy models, which can also be thoroughly understood by conventional means. In addition, a toy model combined with artificially generated data has the benefit of being computationally less expensive to calculate, which allows one to investigate the effects of non-linear parametrisation to a higher level of numerical accuracy.

Consider a dataset $\big\{(x_i, y_i, \sigma_i)\big\}$ where $\dim \mathcal{X} = 1 = \dim \mathcal{Y}$, that is, both the observation conditions x_i and observations y_i have only a single component. The true underlying model from which the data is generated features a quartic relationship between x and y that can be modelled using two positive non-zero constants $\alpha, \beta \in \mathbb{R}^+$ by $y_i(x_i) = \alpha\, x_i^4 + \beta + \epsilon_i$ where ϵ_i presents some measurement error which is distributed according to $\epsilon_i \sim N(0, \sigma_i)$.

Clearly, when attempting to model this dataset, the best results are achieved by a model which encapsulates the quartic relationship which was used to generate the data in the first place. Therefore, one should clearly choose the model function as

$$y_{\text{lin}}(x; \theta) = y_{\text{lin}}(x; \alpha, \beta) = \alpha\, x^4 + \beta \tag{5.1}$$

where the parametrisation $\theta = (\theta^1, \theta^2) = (\alpha, \beta)$ is in some sense the most natural choice since the parameters α and β appear linearly. It is not hard to see that this model is injective with respect to the parameters α and β for all values, wherefore one might consider the maximal model manifold to be $\mathcal{M} \cong_{\text{top.}} (\mathbb{R}^2, \mathcal{O}_{\text{std}})$.

An alternative yet equally valid parametrisation of this quartic relationship is given by

$$y_{\text{non-lin}}(x; \theta) = y_{\text{exp}}(x; \alpha, \beta) = 15\alpha^3\, x^4 + \beta^5. \tag{5.2}$$

Whereas the first model y_{lin} is linear with respect to the fit parameters α and β, the second model $y_{\text{non-lin}}$ clearly is not.[65] The effective parameter transformation from the linear to the non-linear model is given by $(\alpha, \beta) \mapsto (15\alpha^3, \beta^5)$ which can be inverted to yield $(\tilde{\alpha}, \tilde{\beta}) \mapsto \left(\sqrt[3]{\tilde{\alpha}/15}, \sqrt[5]{\tilde{\beta}}\right)$.

However, it is easy to see that the inverse transformation is not differentiable for $\tilde{\alpha} = 0$ or $\tilde{\beta} = 0$. In particular, this means that the inverse transformation cannot be smooth in this regime and therefore

[65] In principle, one could of course exaggerate this comparatively mild non-linearity even further, for example by repeated applications of functions such as $\sinh$ or $\exp$ to the parameters.

the non-linear parametrisation chart must be restricted to the domain $(\mathbb{R}^+ \setminus \{0\}) \times (\mathbb{R}^+ \setminus \{0\})$ to remain compatible with the original parameter manifold. Clearly, the non-linear model is also injective with respect to its parameters on this domain.

Already, this shows that the domain on which the parametrisation of a model is valid crucially depends on the specific choice of parametrisation. It is observed in figures 46 and 47 that such a chart boundary generally reveals itself in practice by the fact that the model becomes locally structurally unidentifiable, i.e. that the geometric density vanishes.

A plot of the aforementioned toy dataset is depicted in figure 40. While the parameters used to generate said dataset are $\theta_{\text{true}} = (1.0, 8.0)$, the maximum likelihood estimate comes out as $\theta_{\text{MLE}} \approx (1.0958, 7.8802)$. This discrepancy is of course a result of the statistical errors ϵ_i which were deliberately introduced as well as the small size of the dataset which contains only 16 data points.

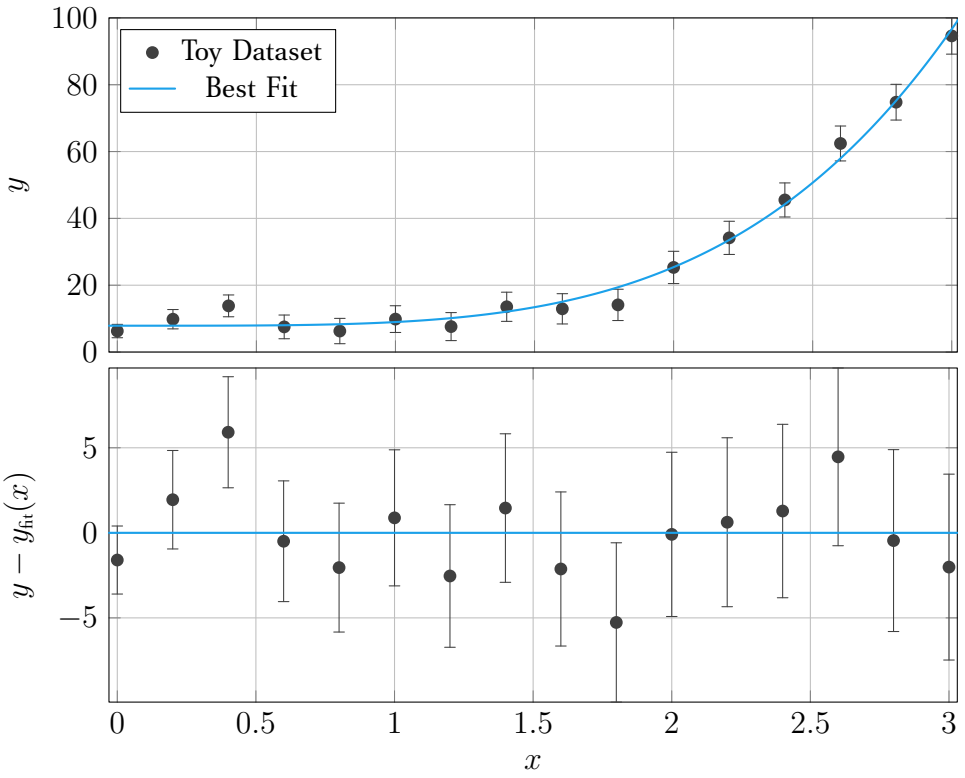


Figure 40: Plot of the quartic toy dataset together with the best fit of approximately $y_{\text{model}}(x, \theta_{\text{MLE}}) \approx 1.0958x^4 + 7.8802$ is shown on top and the corresponding residual plot where the model prediction has been subtracted from all values is shown below. Especially from the residual plot, one can see that the data is described really well by the quartic model. This is of course unsurprising given that a quartic relationship was used to generate the dataset in the first place.

While a discrepancy between the true parameters and the inferred best fit parameters is to be expected, the immediate question is of course what the uncertainty in the maximum likelihood estimate is and whether the true parameters are contained in a confidence region of reasonable

confidence level. As figures 41 and 42 show, the true parameter configuration is well within the 2σ confidence region irrespective of the linearity of the chosen parametrisation. Specifically, the log-likelihood ratio test indicates that the true parameter configuration is situated on the boundary of the confidence region corresponding to the confidence level $q \approx 89.3\,\% \mathrel{\widehat{=}} 1.61\sigma$ in both cases.

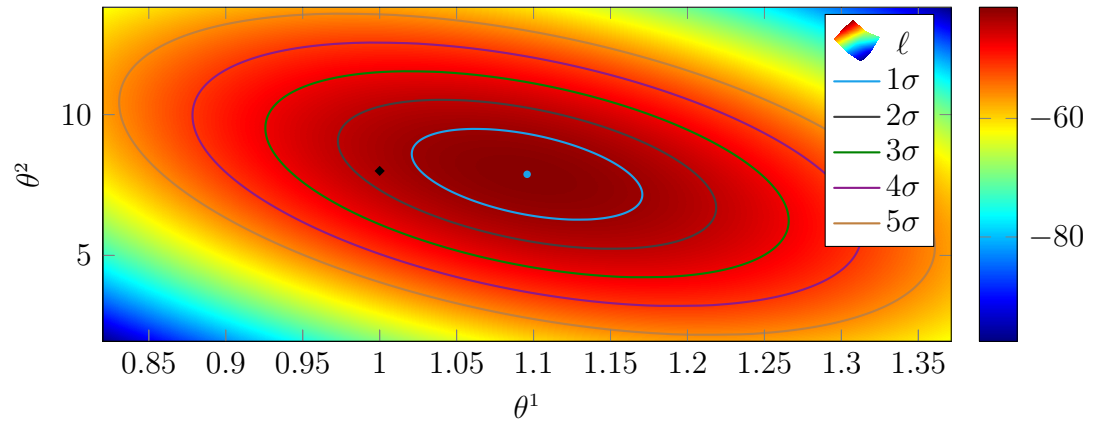


Figure 41: Plot of the log-likelihood function ℓ for the linearly parametrised quartic toy model defined in equation (5.1). This illustration reaffirms that the iso-likelihood curves (i.e. the boundaries of confidence regions) form perfect ellipses around the maximum likelihood estimate, which in this case is approximately located at $\theta_{\text{MLE}} \approx (1.0958, 7.8802)$. The true parameter configuration is indicated by a black diamond which is clearly located inside the 2σ confidence region.

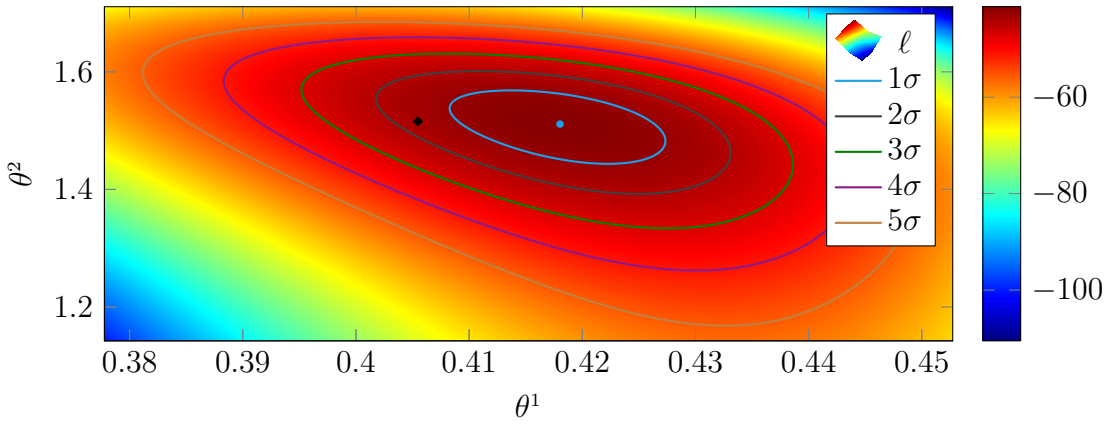


Figure 42: Plot of the log-likelihood function ℓ for the non-linear quartic toy model from equation (5.2) and the associated confidence regions of levels 1σ up to 5σ. One can clearly see that the confidence regions increasingly deviate from an ellipsoidal shape for higher confidence levels. Unsurprisingly, the change in parametrisation also affects the position of the maximum likelihood estimate, which is approximately located at $\theta_{\text{MLE}} \approx (0.4180, 1.5112)$ for the non-linear model. The true parameter configuration is indicated by a black diamond which is clearly located inside the 2σ confidence region exactly as for the linear model in figure 41.

Figure 42 indicates that the shapes of the confidence regions become increasingly irregular for

higher confidence levels as a result of the non-linear parametrisation of the model. Therefore, consideration of the exact shape of the confidence intervals in the chosen parametrisation is especially important when aiming for higher confidence levels. This becomes even more apparent in figure 44, which gives a larger view of the parameter manifold around the MLE.

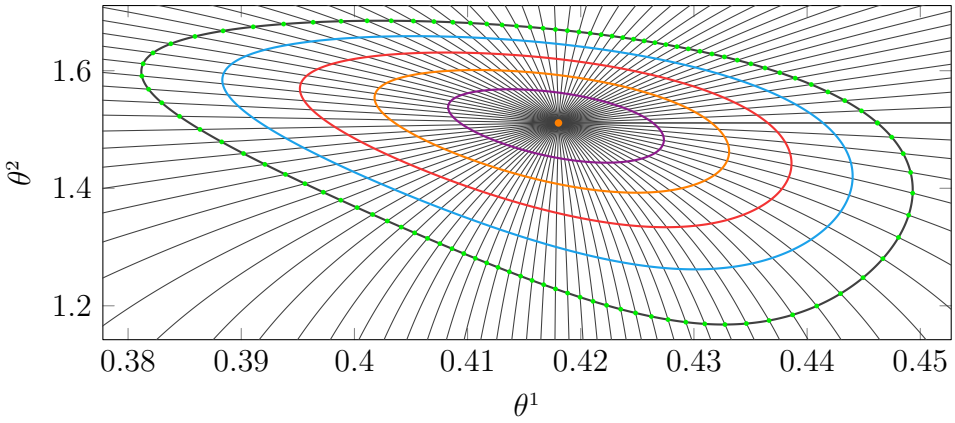


Figure 43: Plot of 100 radial geodesics emerging from the maximum likelihood estimate on $\mathcal{M}$ for the non-linear toy model. The green points indicate where each of the geodesics respectively reaches a length of 5.361, which is visually indistinguishable from the 5σ confidence boundary. Detailed statistics on the particular values of geodesic length at which each of the confidence boundaries are intersected can be found in table 2.

Figure 43 nicely illustrates that the geodesic distance from the MLE to the boundaries of confidence regions is the same in all directions, even in the case of a non-linearly parametrised model. Furthermore, the lengths of said radial geodesics up to their respective intersections with confidence boundaries coincide to the length predicted via equation (4.75) up to the precision with which the geodesics were constructed. The Christoffel symbols for this computation were obtained from the metric via finite difference computations and therefore represent the main bottleneck in accuracy. All in all, this represents compelling evidence in support of the conjecture that geodesic distance provides a highly accurate approximation to the likelihood ratio test for normal likelihoods.

Essentially, the geometric density factor can be interpreted as accounting for shifts in probability volume due to coordinate transformations, since it is designed to produce the inverse factor that is created by transforming the integration measure such that they cancel and leave the overall integral invariant. Therefore, the volume of a confidence interval $\mathcal{C}_q$ of level q, as computed via

$$\mathrm{vol}(\mathcal{C}_q) = \int_{\mathcal{C}_q} \sqrt{\det(g)} = \int_{\theta(\mathcal{C}_q)} \mathrm{d}\theta \, \sqrt{\det\big(g(\theta)\big)} \tag{5.3}$$

must remain invariant under changes of coordinate chart. For models which are linear in their parameters (and locally structurally identifiable), the geometric density is a constant factor on $\mathcal{M}$. As an alternative to direct computation, one can easily infer this from the fact that if the model map y_θ is linear in the parameters, then the push-forward map h_* generated from the embedding

Confidence Level q	Mean Geodesic Distance of Boundary from MLE
1σ	$1.515\,173 \pm 4.2 \cdot 10^{-14}$
2σ	$2.485\,976 \pm 5.6 \cdot 10^{-14}$
3σ	$3.439\,354 \pm 6.7 \cdot 10^{-14}$
4σ	$4.397\,034 \pm 7.7 \cdot 10^{-14}$
5σ	$5.361\,315 \pm 8.6 \cdot 10^{-14}$

Table 2: Table of lengths at which the confidence boundaries of 1σ to 5σ are reached by 1000 geodesics with unique initial starting directions, similar to the illustration in figure 43. The given uncertainties in the mean geodesic distance represent one standard deviation and are limited by the accuracy of the numerically calculated geodesics. The shown geodesic distances coincide with the distance predicted by the quantile function of the χ^2_2 distribution via $\sqrt{F_2^{-1}(q)}$ up to the precision to which the geodesics were computed.

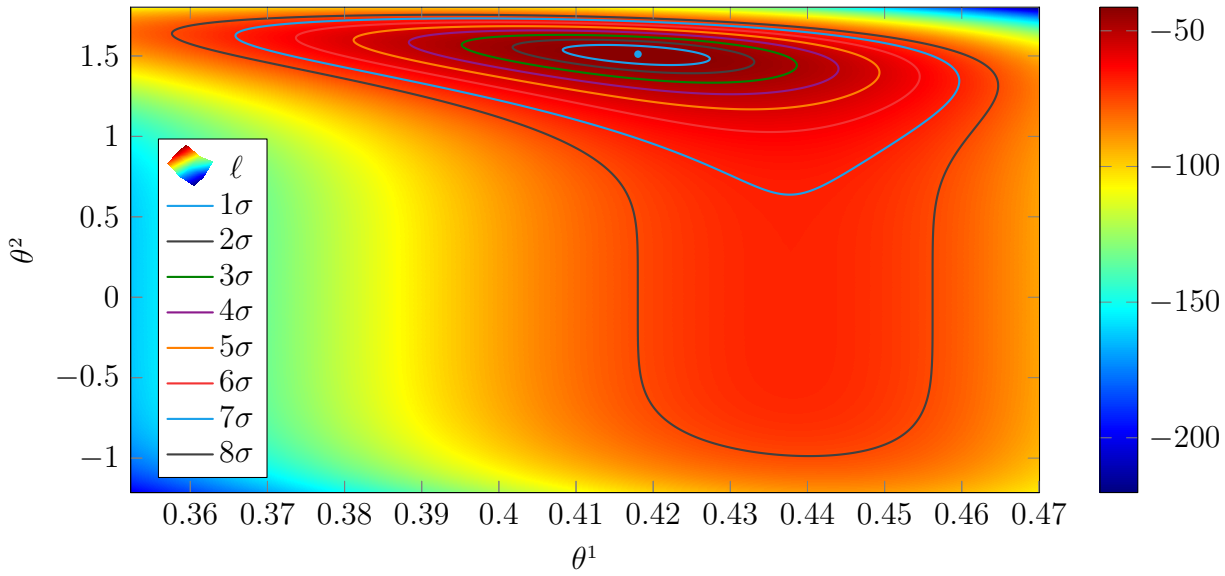


Figure 44: Larger view of the log-likelihood function ℓ for the non-linearly parametrised toy model where the boundaries associated with the confidence regions from 1σ to 8σ are shown. Although it is known in this case that the chart of the non-linearly parametrised toy model is no longer valid if the second parameter $\theta^2 \leq 0$, it is important to note that the log-likelihood function offers no visual indication of this.

map $h(\theta) = \big(y_{\text{model}}(x_1;\theta), ..., y_{\text{model}}(x_N;\theta)\big)$ must necessarily be constant, implying that the pull-back of a constant metric $g_{\mathcal{D}}$ results in a constant metric $g_{\mathcal{M}}$.

This suggests that the volume $\text{vol}(\mathcal{C}_q)$ can potentially be used as a consistent measure of the overall uncertainty in the parameters associated with different confidence regions. However, it is not known at this point whether the volume of confidence regions can be used to compare the descriptions of a dataset through models which differ in the dimensionality of their parameter manifold. That is to say, it should be further investigated whether this measure can be used

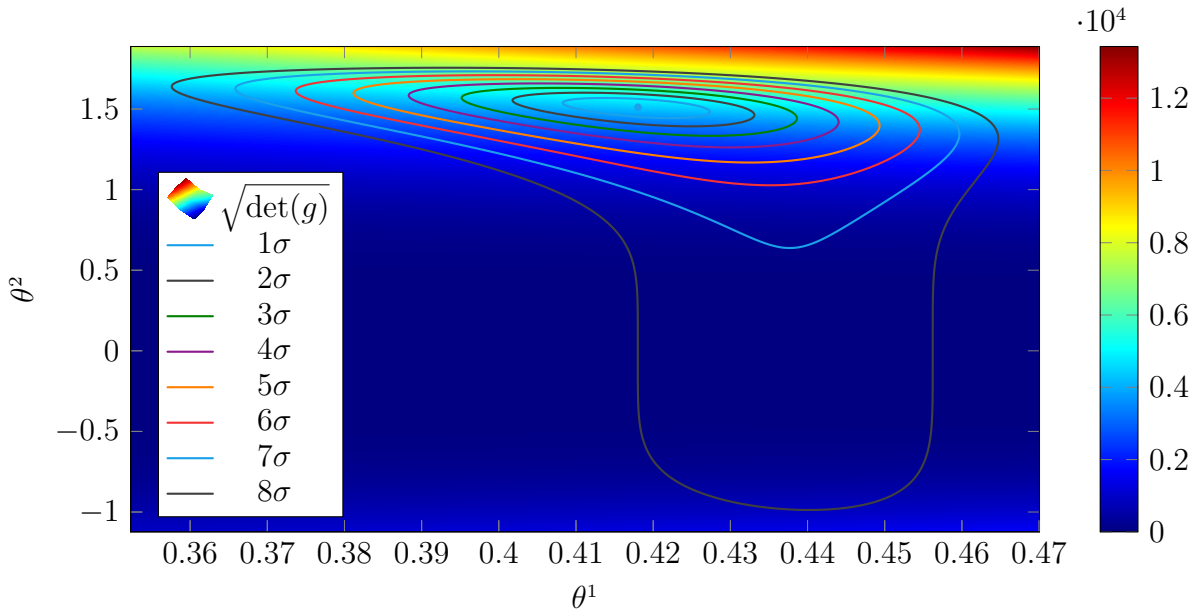


Figure 45: Plot of the geometric density factor $\sqrt{\det(g)}$ on the parameter space $\mathcal{M}$ of the non-linear toy model. The depicted area both contains areas where the geometric density is large, i.e. $\sqrt{\det g} > 1.2 \cdot 10^4$ and also areas where it is very small. As a result, the linear scaling with respect to the geometric density does not constitute a useful representation. Instead, a more appropriate scaling is shown in figure 46.

not only to compare different parametrisations of the same model but to compare models which structurally differ in their predictions.[(66)] In addition, the behaviour of the volume of confidence regions of identical models and parametrisations for different datasets might be instructive. For example, by how much the volume of a confidence region decreases by adding a single further data point depending on its location, i.e. how volume decreases depending on whether the added point is an "outlier".

By imagining a smooth transition from a regime where the model parametrisation is linear and the geometric density is constant on the one hand to a regime where the non-linearity in the model parametrisation is incrementally switched on in a perturbative sense, it is not hard convince oneself that the boundaries of confidence intervals will slowly start to deform in the direction of decreasing geometric density while also balancing the fact that they must contain the maximum likelihood estimate. Thus, one may deduce that, in a sense, the geometric density is responsible for the non-linear shapes of the confidence regions and its gradient can be used to qualitatively surmise in which directions a confidence region will be distorted before specifically calculating its exact boundary.

Moreover, it might be possible to exploit either the geometric density and its gradient or the embedding condition stated in equation (3.33), to construct a chart transformation which compensates for the non-linear distortions induced by a parametrisation and renders the confidence regions

[(66)] For example by choosing model corresponding to a parabolic relationship between the observations $y \in \mathcal{Y}$ and the conditions $x \in \mathcal{X}$ instead of a quartic relationship for the toy dataset.

more spherical in the transformed coordinates.[67]

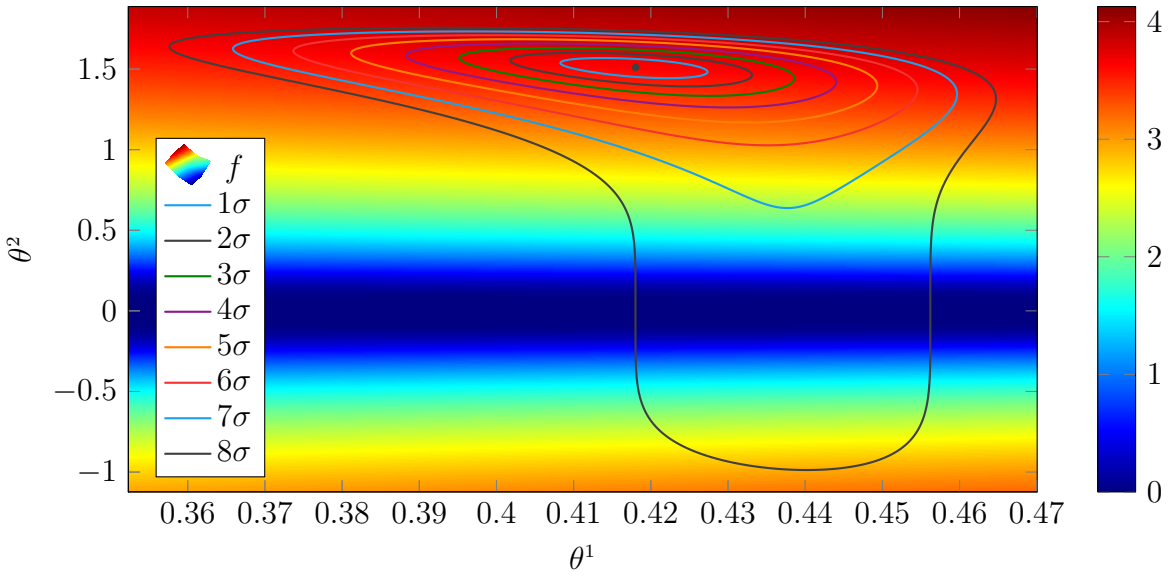


Figure 46: This plot depicts the geometric density factor of the non-linear toy model with a logarithmic rescaling of the form $f = \log_{10}\left(1 + \sqrt{\det(g)}\right)$. As a result, it is significantly easier to visually discern the relevant features of the geometric density, such as the apparent mirror symmetry around the $\theta^1 = 0$ line. Also, the geometric density vanishes precisely on the $\theta^2 = 0$ line.

Given that the geometric density vanishes on the $\theta^2 = 0$ line in the non-linear parametrisation of the toy model as shown in figures 46 and 47 which indicates that the model is no longer structurally identifiable on this line, one should consider this line a boundary of the chart defined by this non-linear parametrisation. For points $\theta^2 \leq 0$, this chart is no longer compatible with the chart induced by the linear parametrisation, since the chart transition map is no longer smooth. As a result, the fact that the non-linearly parametrised toy model is not locally structurally identifiable everywhere along the 8σ confidence region implies that it not longer constitutes a valid confidence region in this choice of coordinates. Instead, another chart which contains the missing "lower part" of this confidence region is needed to provide a consistent representation.[68]

5.2 Analysis of the Distance–Redshift Relationship of Type 1A Supernovæ

The Supernova Cosmology Project (SCP) was founded in 1988 at Berkeley National Laboratory and headed by S. Perlmutter with the goal of using type 1A supernovæ to measure the rate of expansion of the Universe. Since supernovæ of this particular type detonate with a known luminosity, their observed apparent brightness can be used to estimate their distance which is why they are often

[67] Of course, to find a chart in which the confidence regions are perfect hyperellipses requires the metric to be constant, wherefore the parameter manifold must be flat for such a transformation to exist.

[68] Given that this is an artificially constructed model, such a chart is of course induced by the linear parametrisation of the toy model y_{lin}. Hence, it is known that the 8σ region does objectively exist in $\mathcal{M}$. Yet, this may generally not be the case.

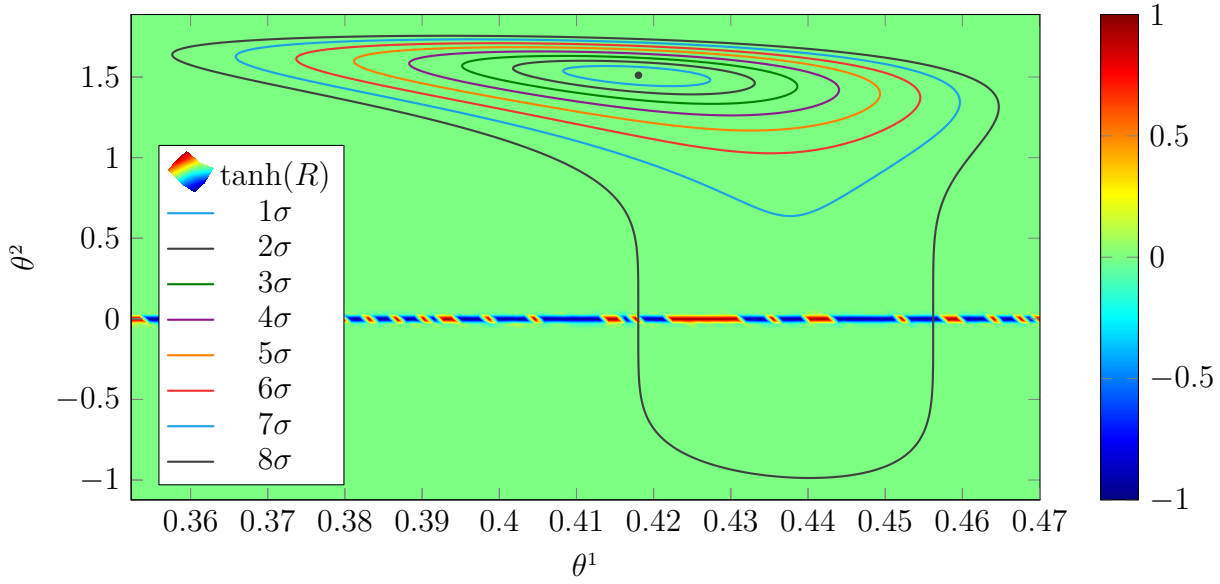


Figure 47: Plot of $\tanh\big(R(\theta)\big)$. Of course, the curvature features are visually over-exaggerated in this depiction as a result of the hyperbolic rescaling of the Ricci scalar. One can see that the Ricci scalar is basically zero everywhere, except around the line defined by $\theta^2 = 0$, on which it strongly fluctuates. As observed in figure 46, the model is not locally structurally identifiable on this line. On balance, one may conclude that the region $\theta^2 \leq 0$ is no longer a valid part of the chart domain and that the parameter manifold is flat.

referred to as standard candles. In 1998, the SCP first published evidence that the expansion of the Universe appears to be accelerating instead of slowing down in collaboration with another group, for which they were jointly awarded the 2011 Nobel prize in Physics.[69] The SCP dataset which is used in the following contains 580 independent measurements of distant type 1A supernovæ and is publicly available via [1]. An analysis of this dataset by conventional methods can be found for example in [74].

One possible way of quantifying distance in a cosmological setting is via the so-called distance modulus μ which can be expressed as a function of the cosmological redshift z by

$$\mu(z;\Omega_{\mathrm{m}},w) = 10 + 5\log_{10}\left((1+z)\,d_{\mathrm{H}}\int_0^z \mathrm{d}x\;\frac{1}{\sqrt{\Omega_{\mathrm{m}}\,(1+x)^3 + (1-\Omega_{\mathrm{m}})(1+x)^{3(1+w)}}}\right) \tag{5.4}$$

where $d_{\mathrm{H}} = c/H_0$ is the Hubble distance, Ω_{m} is the matter density in the Universe as observed today and w is the dark energy equation of state. As indicated by the notation $\mu(z;\Omega_{\mathrm{m}},w)$, the model is interpreted as having only two variable parameters Ω_{m} and w, whereas Hubble's constant is already fixed at an assumed value of $H_0 = 70\,\frac{\mathrm{km}}{\mathrm{s\cdot Mpc}}$. While it would technically be possible to also infer the value of H_0 and other parameters from the SCP dataset, its relatively small size

[69] Very recently, the reliability of type 1A supernovæ as standard candles has come into question as observations indicate that there may be mechanisms for stars which were previously theorised to always detonate at the predictable luminosity which is characteristic for type 1A supernovæ to undergo only partial detonations (see [25]).

means that the number of parameters that can be simultaneously inferred from it to a meaningful level of accuracy is fairly limited.

As a result, this distance modulus model implicitly assumes that the Universe is filled with only two types of cosmological fluids: on the one hand, a matter fluid with density parameter Ω_m and an associated equation of state parameter $w_\mathrm{m} = 0$ which accounts for the effects of both baryonic matter as well as dark matter. On the other hand, it also describes a fluid with density Ω_Λ which models the effects of dark energy. Since the Friedmann equations require that the sum of all density parameters is normalised to one, the density parameter of the dark energy fluid must be given by $\Omega_\Lambda = (1 - \Omega_\mathrm{m})$ under the assumption that the Universe is flat.[70] However, the equation of state parameter w associated with dark energy must be inferred from the data. Another assumption of this model is that the effects of radiation, which dominated the expansion behaviour of the Universe at early cosmological times are negligible at late cosmological times (i.e. today), where instead the effects of matter and dark energy dominate.

Given that a definite integral can be reformulated as an ordinary differential equation of the form

$$F(z) = \int_a^z \mathrm{d}x\, f(x) \quad \Longleftrightarrow \quad \frac{\mathrm{d}F}{\mathrm{d}z}(z) = f(z) \quad \text{with the initial condition} \quad F(a) = 0, \tag{5.5}$$

this allows one to make use of sophisticated high-order Runge–Kutta integration algorithms that have been developed for ODEs in the numerical evaluation of such integrals. Fortunately, numerical integration can even be entirely avoided in this particular case since there exists a closed form solution for the integral in equation (5.4) which can be found by symbolic integration algorithms to be

$$\int \mathrm{d}u\, \frac{1}{\sqrt{A\,u^3 + B\,u^c}} = -\frac{2u\sqrt{\frac{A\,u^{3-c}}{B} + 1}\cdot {}_2F_1\left(\frac{1}{2}, \frac{c-2}{2c-6}; \frac{3c-8}{2c-6}; -\frac{A\,u^{3-c}}{B}\right)}{(c-2)\sqrt{A\,u^3 + B\,u^c}} + \text{const.} \tag{5.6}$$

where ${}_2F_1$ is the hypergeometric function defined by

$${}_2F_1(a, b; c; z) := \sum_{k=0}^{\infty} \frac{\Gamma(a+k)\,\Gamma(b+k)\,\Gamma(c)}{\Gamma(a)\,\Gamma(b)\,\Gamma(c+k)}\,\frac{z^k}{k!} \tag{5.7}$$

$$= \frac{\Gamma(c)}{\Gamma(b)\,\Gamma(c-b)} \int_0^1 \mathrm{d}t\, t^{b-1}\,(1-t)^{c-b-1}\,(1-zt)^{-a}. \tag{5.8}$$

Thus, existing approximations of the hypergeometric function ${}_2F_1$ can be used to efficiently compute solutions to the definite integral in equation (5.4) which not only decreases the overall computational effort significantly, but also increases the accuracy compared with numerical integration schemes.

To ensure that the distance modulus $\mu(z; \Omega_\mathrm{m}, w)$ constitutes a globally structurally identifiable model (see section 3.6), it must be checked whether it is injective with respect to the parameters

[70] Curvature can also be treated as a fluid in cosmological spacetimes using a density parameter Ω_K which, however, is assumed to be zero here. Hence, the Friedmann equations dictate $\Omega_\mathrm{m} + \Omega_\Lambda \overset{!}{=} 1$.

Ω_m and w. Specifically, one must show that

$$\mu(z;\Omega_\mathrm{m},w)=\mu(z;\tilde{\Omega}_\mathrm{m},\tilde{w})\quad\forall z\in\mathbb{R}_0^+\qquad\Longrightarrow\qquad\Omega_\mathrm{m}=\tilde{\Omega}_\mathrm{m}\quad\wedge\quad w=\tilde{w}. \tag{5.9}$$

Indeed, by exploiting the differentiability of $\mu(z;\Omega_\mathrm{m},w)$ with respect to the redshift z, it is straightforward to demonstrate that the model $\mu(z;\Omega_\mathrm{m},w)$ is injective as long as $w\neq 0$ and $\Omega_\mathrm{m}\neq 1$.

Mainly for physical reasons, it is assumed that the matter density Ω_m is larger than zero. This is also sensible from a mathematical point of view since a negative value for the matter density parameter can cause the square root in the model to become complex-valued: concretely, if $w=-1$, then the square root in the model μ becomes

$$\sqrt{\Omega_\mathrm{m}\,(1+z)^3+(1-\Omega_\mathrm{m})(1+z)^{3(1+w)}}\overset{w=-1}{=}\sqrt{\Omega_\mathrm{m}\,(1+z)^3+(1-\Omega_\mathrm{m})} \tag{5.10}$$

which yields complex values for $\Omega_\mathrm{m}<0$ and sufficiently large redshifts $z\gg 1$. Hence, for $\mu(z;\Omega_\mathrm{m},w)$ to retain its predictivity at any redshift $z\in\mathbb{R}_0^+$, values of $\Omega_\mathrm{m}<0$ should be excluded.

In summary, the maximal open connected domain of admissible parameter configurations on which the model is injective is given by

$$\mathcal{M}=\left\{(\Omega_\mathrm{m},w)\in\mathbb{R}^2\;\middle|\;0<\Omega_\mathrm{m}<1,w<0\right\}=(0,1)\times(\mathbb{R}^-\setminus\{0\}). \tag{5.11}$$

Although such a consideration of the maximal injective domain of a model in no way excludes the possibility that the true value of the parameters may indeed lie outside of the injective region (e.g. $w\geq 0$ or $\Omega_\mathrm{m}\leq 0$), it nevertheless indicates that the chosen model is not suited to describing these regimes.

The SCP dataset is illustrated in figure 48 together with the distance modulus from equation (5.4) evaluated at the maximum likelihood estimate $(\Omega_\mathrm{m},w)\approx(0.28,-1.00)$. Moreover, the corresponding exact confidence regions of levels 1σ and 2σ are shown together with the log-likelihood function in figure 49. Especially from figure 50, it is evident that iso-likelihood contours are of non-ellipsoidal shape which demonstrates the non-linear dependence of the model and in turn the likelihood on the model parameters Ω_m and w.

While it is in principle possible to construct confidence regions whose level is arbitrarily close to one, the maximum likelihood estimate of the SCP dataset is relatively close to the established boundary of the parameter manifold. Consequently, for some confidence level q', the confidence region will make contact with the boundary of $\mathcal{M}$. For any confidence levels $q>q'$ the interpretation of the confidence region becomes tricky: given that a part of the confidence region lies "outside" of $\mathcal{M}$, the confidence level q loses its quantitative meaning if this is not accounted for.

By performing a parameter search on the line defined by $\Omega_\mathrm{m}=10^{-15}$, one finds that the likelihood

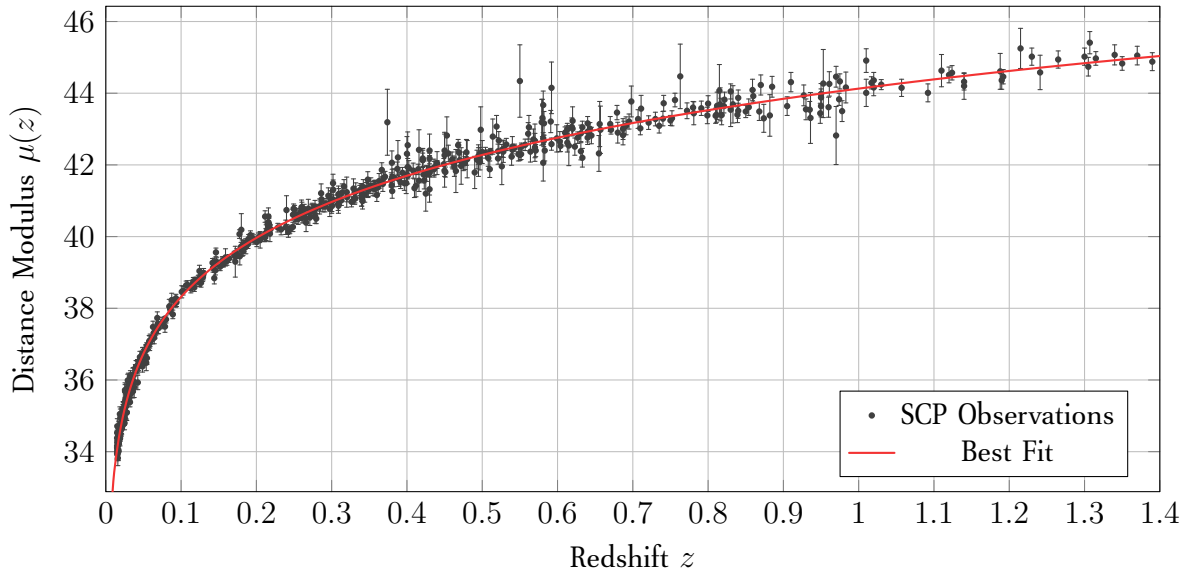


Figure 48: Visualisation of the distance modulus $\mu(z)$ as a function of redshift z for observations of type 1A supernovæ, measured by the Supernova Cosmology Project. A fit of equation (5.4) to the data was performed using maximum likelihood estimation to determine the optimal parameters of $\Omega_\mathrm{m} \approx 0.28$ and $w \approx -1.00$ under the assumption of $H_0 = 70\,\frac{\mathrm{km}}{\mathrm{s{\cdot}Mpc}}$. The data was excerpted from [1].

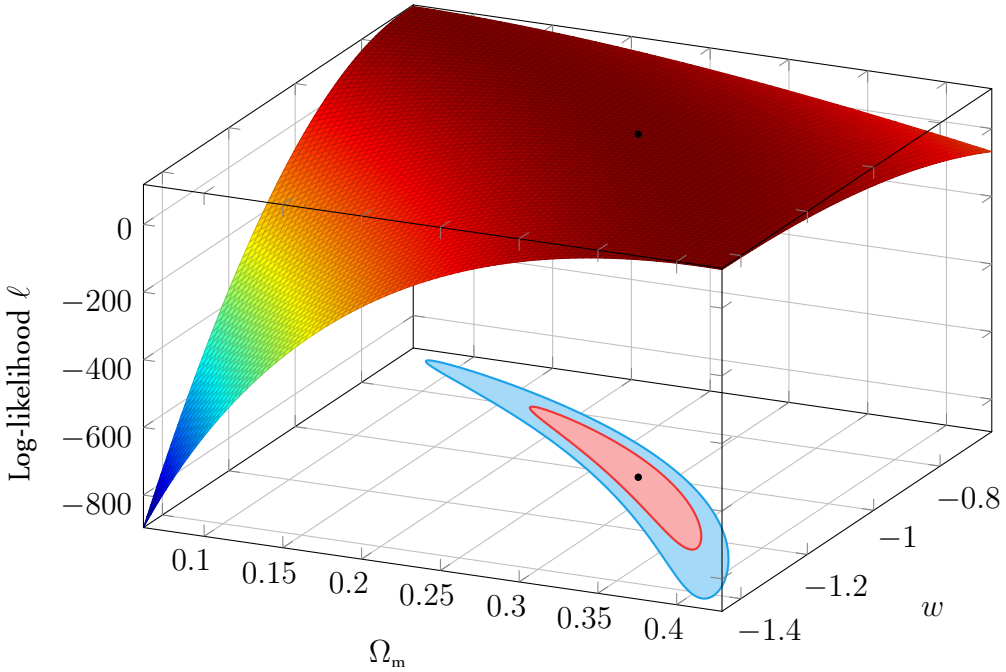


Figure 49: Surface plot of the log-likelihood $\ell\big(\mathrm{data} \,\big|\, (\Omega_\mathrm{m}, w)\big)$ as a function of the parameters Ω_m and w around the maximum likelihood estimate. Unsurprisingly, the confidence regions of levels 1σ and 2σ closely follow the ridge of the likelihood surface.

attains its maximum at $w \approx -0.6072$. Computation of the likelihood ratio test reveals

$$\lambda := \ell(\mathrm{data} \,|\, \theta_\mathrm{MLE}) - \ell\big(\mathrm{data} \,\big|\, (10^{-15}, -0.6072)\big) = \frac{1}{2}\, F_2^{-1}(q) \tag{5.12}$$

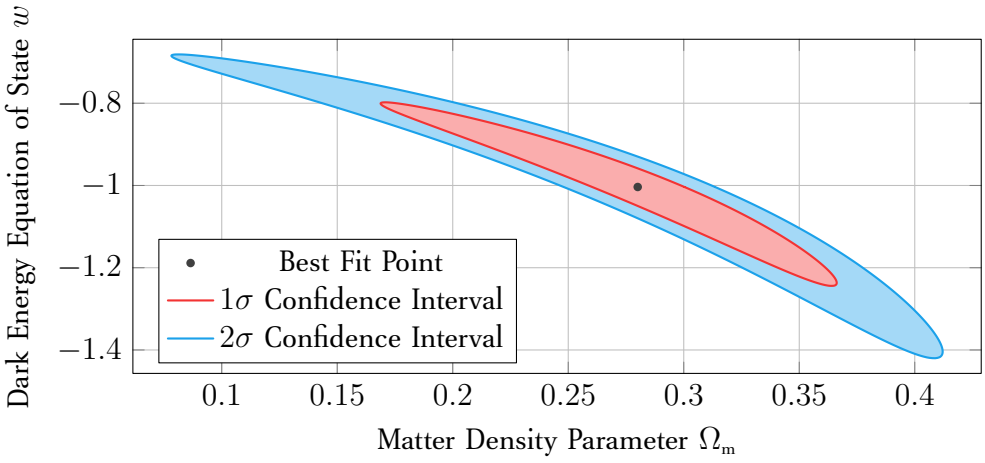


Figure 50: Plot of the exact 1σ and 2σ confidence regions around the maximum likelihood estimate as generated by the geometric iso-likelihood method from section 4.3 for the SCP dataset. The non-linear dependence of the model function on the parameters Ω_m and w can be clearly seen from the distorted shape of the confidence regions. Their respective volumes can computed as approximately $\mathrm{vol}(\mathcal{C}_{1\sigma}) = 51.4 \pm 0.1$ and $\mathrm{vol}(\mathcal{C}_{2\sigma}) = 492 \pm 20$.

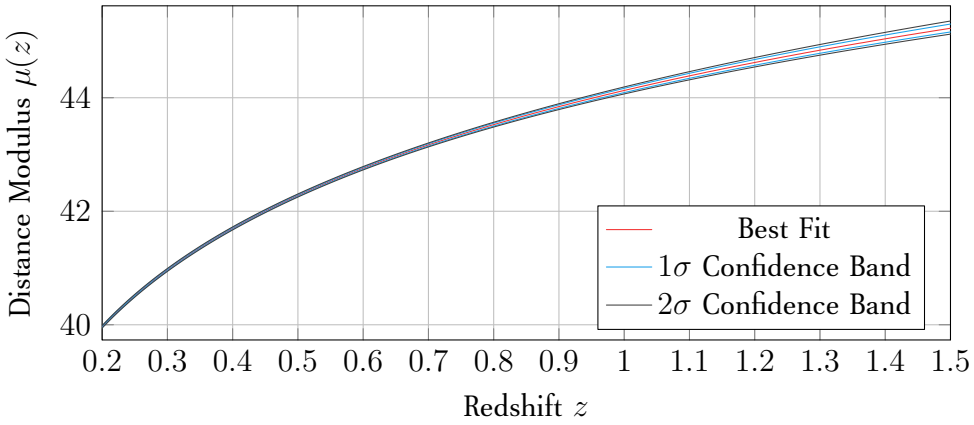


Figure 51: Plot of pointwise confidence bands for the SCP dataset and distance modulus model around the maximum likelihood prediction using the method described in section 4.5. Evidently, the model prediction is strongly constrained by the available data for small and medium redshifts z whereas uncertainty in the model prediction increases for higher redshift. This is an indication that further observations at high redshift would decrease the remaining uncertainty in the parameters and therefore constrain the model prediction the most.

$$\iff F_2(2\lambda) = F_2\Big(F_2^{-1}(q)\Big) = q = \operatorname{erf}\left(\frac{n}{\sqrt{2}}\right) \tag{5.13}$$

$$\iff \sqrt{2}\,\operatorname{erf}^{-1}\Big(F_2(2\lambda)\Big) = n \qquad \Longrightarrow \qquad n \approx 2.7341 \tag{5.14}$$

where the degrees of freedom between the two hypotheses is $k = \dim \mathcal{M} = 2$. Hence, the largest confidence region possible in this parametrisation which is not connected to the boundary of the manifold $\partial\mathcal{M}$ corresponds to a confidence level of roughly $q \approx 99.4\,\% \mathrel{\widehat{=}} 2.73\sigma$. In the

one-dimensional case, methods to account for a bounded parameter have been proposed e.g. in [85]. In real-world applications such as this, it would be especially helpful to have an extension of such a method with which to account for the boundedness of the admissible parameter domain if it is not possible to find a complementary chart which covers the desired domain.

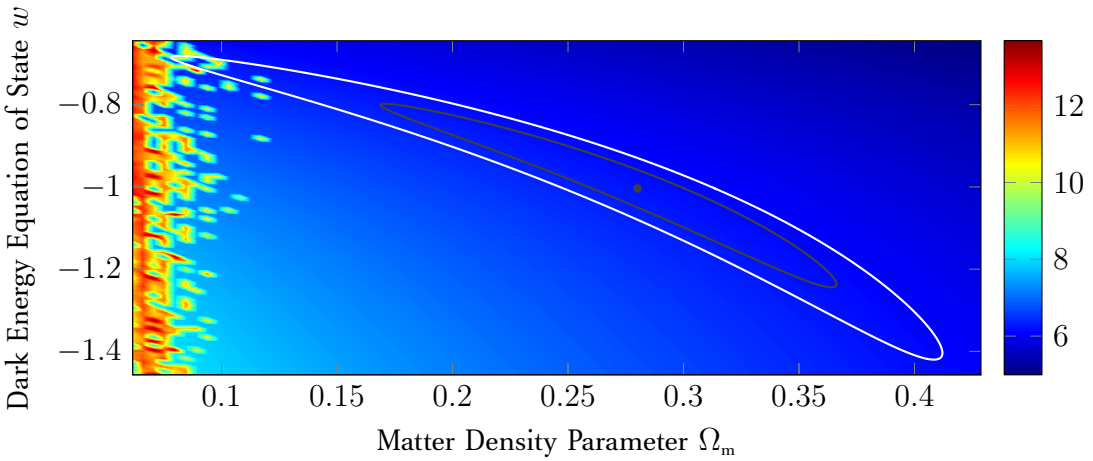


Figure 52: Plot of logarithmic geometric density $\ln\left(\sqrt{\det g}\right)$ for the SCP dataset. A recognisable trend for the geometric density is that it generally decreases for larger values of the matter density Ω_m and larger values of the dark energy equation of state parameter w. One can see that the eigenvalues become very large for small Ω_m wherefore a small change in its value can be seen as specifying a large amount of information.

Since the curvature becomes more and more erratic as one comes closer to the $\Omega_m = 0$ boundary on the parameter manifold $\mathcal{M}$, the ODE solver for confidence boundary generation increasingly runs into instabilities. Therefore, while boundaries of the parameter manifold established in equation (5.11) constitute a conceptual upper limit to for the level of confidence regions that can be established, the instabilities arising from the curvature and geometric density effects represent a practical upper limit to the maximal confidence region which can be computed. Essentially, the construction of confidence regions larger than roughly 2.5σ becomes practically infeasible using the integral manifold method due to the fact that the model is no longer well-behaved in this case. Conferring with the investigation of the non-linearly parametrised toy model (see figure 47), the curvature fluctuations might be interpreted as evidence that the distance modulus model is ill-suited to describe regimes where $\Omega_m \lesssim 0.1$.

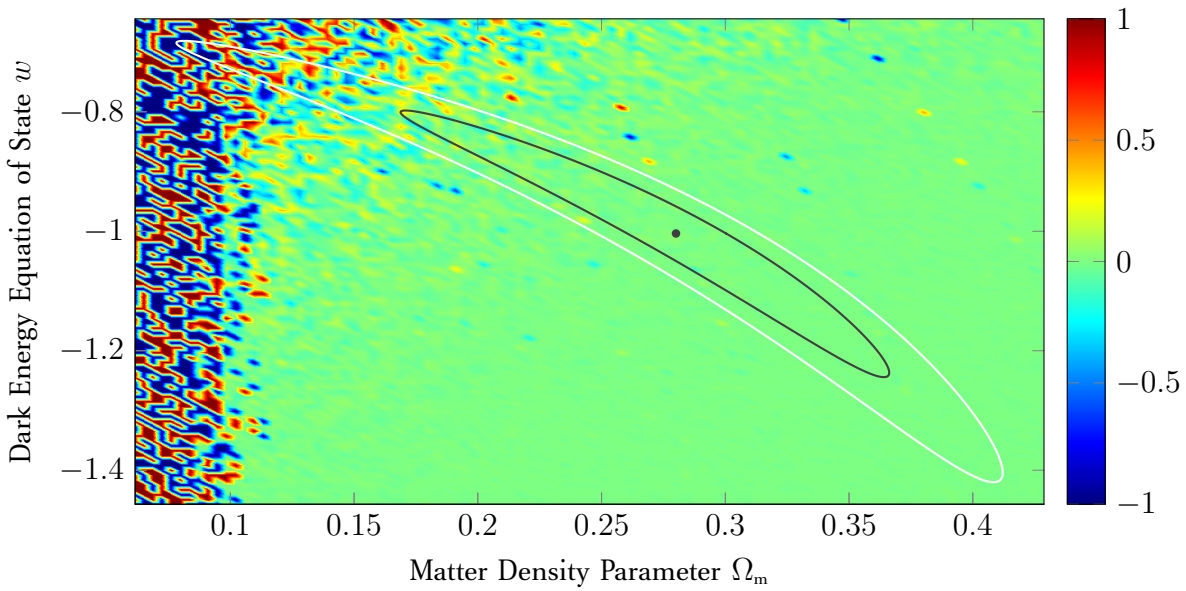


Figure 53: Plot of the Ricci curvature on the parameter manifold associated with the distance modulus model as applied to the SCP dataset after a rescaling according to $\tanh\big(8\tanh(R)\big)$. Using this non-linear rescaling, the relevant features given by large fluctuations in the curvature in regions where $\Omega_m \lesssim 0.1$ which are closely connected to the fluctuations in geometric density. Detailed examination of these regions reveal that the absolute values of some components of the Fisher metric rapidly increase to $\gtrsim 10^7$ wherefore small changes in position strongly affect the prediction.

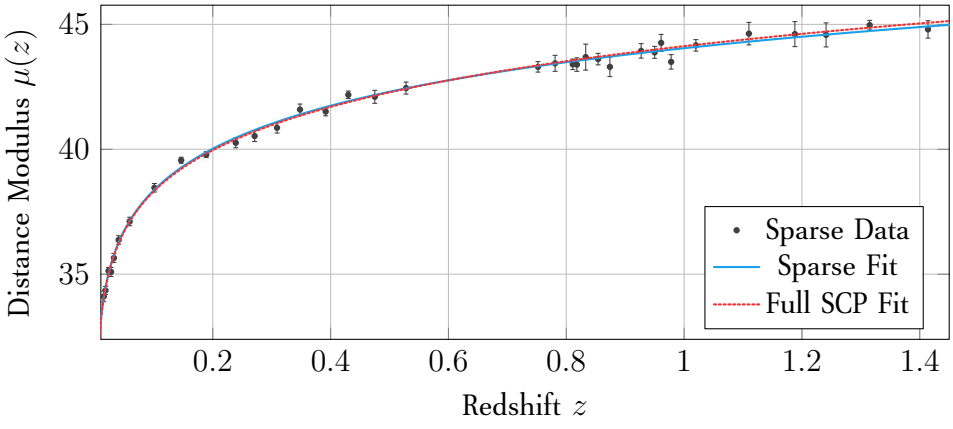


Figure 54: Sparse excerpt of the original SCP dataset containing only $35/580 \approx 6\,\%$ of the data points. This subset of data points was chosen in such a way that representatives for different redshifts z are retained somewhat evenly. The new MLE for the parameters given this sparser dataset is $\theta_{\text{MLE}} \approx (0.506, -2.228)$. Although there is clearly a large discrepancy compared with the MLE obtained for the full dataset, this is unsurprising given the extreme reduction in the number data points.

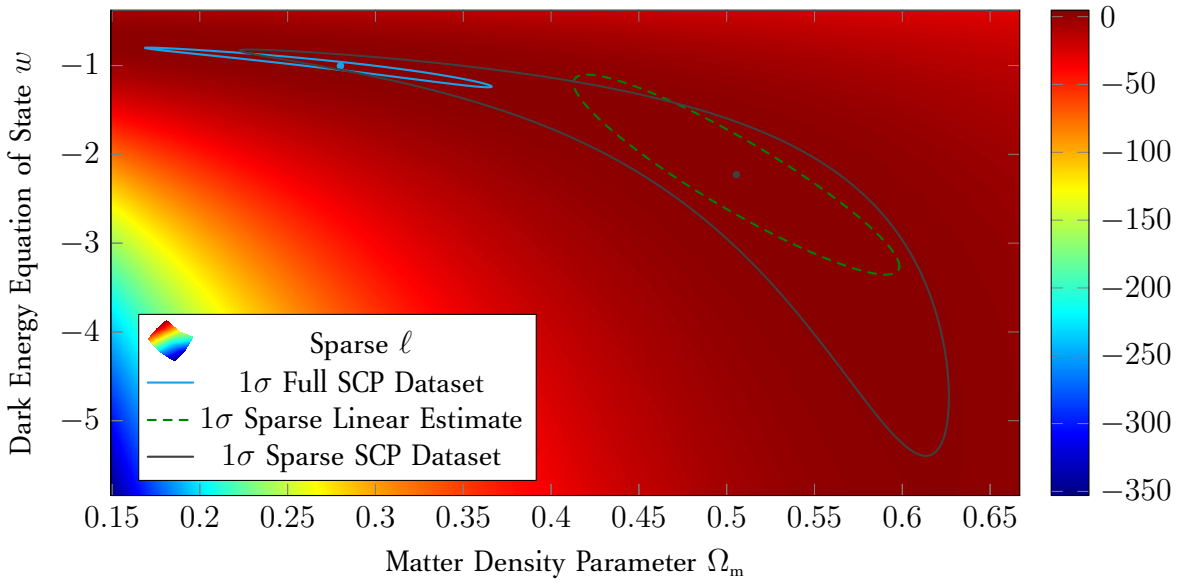


Figure 55: Comparison of the respective locations of the maximum likelihood estimates for the full and sparse SCP datasets and their corresponding exact 1σ confidence regions. Reassuringly, the 1σ confidence region for the sparse dataset still contains the MLE of the full dataset. Integration of the geometric density over the confidence regions yields approximate volumes of 70.1 for the sparse dataset and 50.5 for the full dataset. The dashed ellipse constitutes the Cramér-Rao lower bound estimate for the covariance between the parameters (i.e. the approximate 1σ region) as obtained from the inverse Fisher metric at the MLE. Evidently, this linear estimate grossly misrepresents the true uncertainty in the parameters.

5.3 Performance and Complexity of Schemes for Estimation of Confidence Regions

Given the undeniable conceptual advantages of exact confidence regions over approximative methods, this section aims to compare the performance of the integral manifold method proposed in section 4.3 against other means of establishing exact confidence regions. Of course, the absolute performance of any scheme depends not only on the particular hardware that is used but also on the programming language in which it is implemented. Nevertheless, its general scaling behaviour can provide valuable insight into the efficacy of a method.

Apart from the integral manifold method, there is essentially only one alternative when it comes to computing exact confidence regions which consists of sampling the log-likelihood on a (possibly irregular or non-uniform) grid of parameter configurations. To increase the resulting accuracy, interpolation is used to obtain a piecewise polynomial approximation of the log-likelihood which is typically much more inexpensive to evaluate from a numerical perspective. Finally, the polynomial approximations are used to estimate the location of the intermediate crossings between the sampled grid points where the likelihood takes the desired value which defines the boundary of the confidence region. Clearly, the inherent disadvantage of this scheme is that the overwhelming majority of grid points where the likelihood is sampled are nowhere close to the level set in question, leading to a tremendous waste of computational resources. This disadvantage is further exacerbated for higher-dimensional parameter manifolds.

In an effort to make the comparison the integral manifold method and the interpolating scheme more straightforward, a few simplifying assumptions are made: Especially for large datasets, the evaluation of the log-likelihood constitutes the computationally most expensive part of any scheme for the construction of exact confidence regions.[71] Thus, the main focus is on the necessary number evaluations of the log-likelihood (and other quantities derived from it) whereas ancillary calculations in the compared schemes such as evaluations of polynomial interpolations or root-finding steps are negligible. Likewise, although it is possible in both cases to employ adaptive methods resulting in evaluations of the log-likelihood which are irregularly spaced over the domain, it is assumed that the evaluations of the likelihood are uniformly spaced for either scheme.

It is straightforward to see that in the example of a two-dimensional globally structurally identifiable model, every confidence boundary is topologically equivalent to a circle around the maximum likelihood estimate. To parametrise said boundary as an integral curve to a likelihood-annihilating vector field, the score must be calculated at every point where said vector field is to be evaluated. A single calculation of the components of the score takes on the order of $N \cdot \dim \mathcal{M}$ steps for N data points. Furthermore, this needs to be repeated M times along the topological circle, resulting it roughly $N \cdot \dim \mathcal{M} \cdot M$ evaluations.

In comparison, the grid method for a two-dimensional parameter space takes on the order of N steps per evaluation of the log-likelihood, which must be calculated on a grid of $M \times M = M^{\dim \mathcal{M}}$

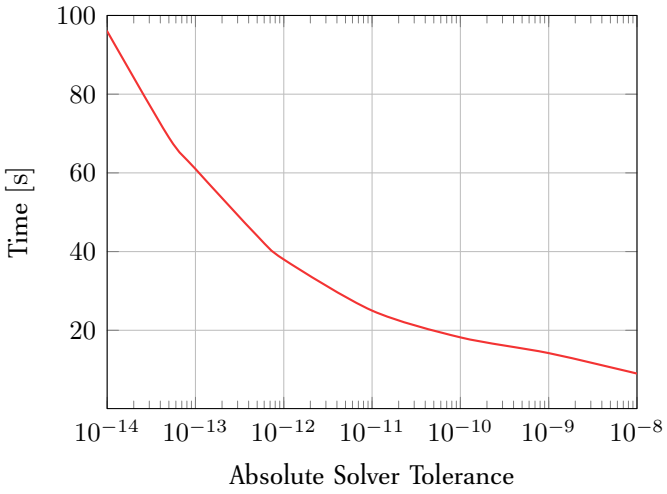
[71] More generally, if the adopted definition of confidence regions is based on any other hypothesis test instead of the likelihood ratio, the same argument applies.

uniformly spaced points (although this may be a different value of M). Already, the overall complexity of the calculation is on the order of $N \cdot M^2$. Thus, one finds in general that the grid sampling involved in the interpolation scheme requires on the order of $O(N \cdot M^{\dim \mathcal{M}})$ evaluations of the log-likelihood whereas the integral manifold method only necessitates $O(N \cdot \dim \mathcal{M} \cdot M^{\dim \mathcal{M}-1})$ evaluations which suggests that the integral manifold method will generally outperform grid sampling methods.

In hindsight, this relationship in the scaling behaviours of both methods is perhaps unsurprising, given that Stokes' theorem $\int_U \mathrm{d}\omega = \int_{\partial U} \omega$ reveals that the operation of taking the topological boundary of a set is intimately connected to the derivative operator. In other words, since the integral manifold method only samples the boundary of the confidence region, its scaling behaviour $O(\dim \mathcal{M} \cdot M^{\dim \mathcal{M}-1})$ essentially corresponds to the derivative with respect to M of the scaling behaviour $O(M^{\dim \mathcal{M}})$ of the grid method, which samples the entire region.

Solver Tolerance	Function Evaluations	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

(a) Table of performance of confidence boundary generation scheme.



(b) Sketch of benchmark performance of the integral curve generation scheme in two dimensions.

Figure 56: Performance of the integral curve scheme for the determination of the 1σ confidence boundary on the SCP dataset, which contains 580 observations. The evaluation of the log-likelihood for this dataset was measured as 0.88 ms while the score was measured as taking 8.8 ms on average. All calculations were executed in a single core computation and using the Tsitouras 5^{th} order Runge–Kutta algorithm.

Figure 56 roughly sketches the performance of the integral manifold method in the specific case of the SCP dataset from section 5.2. From the values listed in the table one finds that the employed solver algorithm required approximately 8000 evaluations of the log-likelihood to achieve an absolute error tolerance of 10^{-14} for the 1σ confidence boundary of the SCP dataset.

Having already established the boundary of the 1σ confidence region, one finds that its dimensions in coordinates are roughly 0.197×0.446. Ignoring the fact that the size and density of the an appropriate sampling grid must usually be determined via trial and error due to the potentially distorted shape of the confidence regions, a total number of 8000 evaluations on a uniform grid of these dimensions results in a grid spacing of approximately $2.1 \cdot 10^{-3}$ in Ω_m and roughly $4.9 \cdot 10^{-3}$

for the parameter w. Even with the assumption that a polynomial approximation between grid points yields results accurate to one hundredth of the width of the grid spacing, this is still orders of magnitude worse when compared with the integral manifold scheme.

In summary, the integral manifold method is significantly more efficient and as a result also more accurate than alternative interpolative methods, especially for higher-dimensional parameter spaces $\mathcal{M}$. It also has the benefit of allowing for more fine-grained control over the desired accuracy to which the confidence region is evaluated through the ability of specifying it to the solver algorithm. On the other hand, it was observed in figure 53 that curvature fluctuations caused by the vicinity to the boundaries of a chart can make the integral manifold method practically infeasible due to numerical instabilities in the integration process. In such cases, grid evaluations of the log-likelihood and subsequent interpolation may be the only viable option for estimating the exact confidence boundary despite its inefficiency.

6 Conclusions

6.1 Summary

In section 2, important notions from the subjects of topology, differential geometry, probability theory and information theory were summarised in so far as they are relevant to later discussions. In particular, the central concepts of coordinate transformations, tensor fields, embeddings, curvature and information divergences were recalled. Next, sections 3.1 to 3.3 defined the notions of datasets, models and introduced the formalism of maximum likelihood estimation.

After the conceptual relationship of the Fisher metric to information divergences was already mentioned in section 2.15, it was then shown in section 3.4 how the Fisher metric can be calculated explicitly when given a dataset and model. In particular, section 3.4.1 highlighted the derivation of the well-known expression for the Fisher metric in the case of normal likelihoods. Subsequently, the central idea of embedding the parameter manifold into the data space as used by Transtrum et al. for example in [78] was presented in section 3.5.

Section 3.6 introduced different concepts of parameter identifiability which are used to judge the suitability of a model. In particular, the injectivity of a model with respect to its parameters was identified as a desirable property, which is referred to as global structural identifiability.

With this in mind, it was explored in section 3.7 how this embedding picture, which was originally only formulated for mutually independent observations with normal error distributions, could be extended to more general error distributions. It was proposed that the Kullback–Leibler divergence should not only be thought of as introducing the Fisher metric on $\mathcal{M}$, but instead as establishing a metric on the data space $\mathcal{D}$ such that the Fisher metric on $\mathcal{M}$ follows from the pull-back under the embedding map. This corresponded to a slight shift in the hierarchy of importance between the Kullback–Leibler divergence and embedding map, since the embedding criterion is thereby taken as the fundamental definition of the relationship between $\mathcal{M}$ and $\mathcal{D}$ instead of the Kullback–Leibler divergence.

This amendment opens up the possibility of exhibiting curvature effects on $\mathcal{D}$ in principle, instead of their being limited to the embedded parameter manifold. Specifically, it was observed in section 3.8.2 that when the moments of an error distribution are directly linked, the natural coordinatisation on $\mathcal{D}$ can be distorted through a non-constant metric on $\mathcal{D}$ even if this ultimately may not result in curvature.

Additionally, it was shown in section 3.8.1 that the expression for the metric on $\mathcal{D}$ obtained for Cauchy error distributions using this principle would result in an additional factor one half compared with the expression one finds for the normal distribution. Further, it was conjectured that the Fisher information associated with a change in the location parameter of any student's t-distribution is bounded by $s^{-2}/2 \leq g_{11}(\mu, s, \nu) \leq s^{-2}$.

Section 4 proceeded by recalling the most widely-used definitions for confidence regions via hypothesis tests. Due to its wide applicability and the Neyman–Pearson lemma, the subsequent discussion focussed in particular on confidence regions defined via the likelihood ratio test.

In anticipation of later results, section 4.2 compared exact confidence regions against various approximations of the uncertainties in the parameters. Among these were the linear approximation of the parameter covariance by inversion of the Fisher information at the MLE, which poses a lower bound on the true covariance according to the Cramér–Rao theorem. In addition, higher order approximations to the likelihood such as DALI—while an improvement over the linear approximation obtained from the Cramér–Rao lower bound—were seen to still produce an inferior representation of the true uncertainty.

One of the main results of this thesis was detailed in section 4.3 and consisted of a general method with which the level sets of a smooth function can be parametrised efficiently. The idea consisted of establishing a Lie algebra of vector fields, whose integral manifolds constitute the level sets of the given function and foliate the parameter manifold. A proof to this effect was given in section 4.3.1 from which it could also be seen, that a sufficient criterion for the existence of such a Lie algebra is that the components of the gradient of said function vanish only at the MLE. In the case of the log-likelihood function, this criterion is equivalent to the non-degeneracy of the Fisher metric since it can be expressed as a product of the first derivatives of the log-likelihood. Also, this requirement is satisfied if the model is globally structurally identifiable.

Section 4.5 discussed how the parametrisation of the exact confidence boundaries can further be used to provide bounds for the uncertainty in the predictions of a model given the uncertainty in its parameter configuration. It was proven in the appendix that it suffices to evaluate the model prediction on the boundary of a confidence region to obtain confidence bands if the model is injective, continuous and the confidence region bounded. This constitutes a significant reduction in the computational effort which is necessary to construct exact (pointwise) confidence bands for the predictions.

Lastly, it was hypothesised in section 4.8 that geodesic distance on the parameter manifold yields an accurate approximation to the likelihood ratio test particularly for normal likelihoods. While the idea of using geodesic distance as a measure of separation between probability distributions was already put forward by C. Rao in 1945, I was unable to find any published results on the direct quantitative relationship between geodesic distance and the likelihood ratio test or the construction of confidence regions despite my best efforts. It was shown that the geodesic distance to any given confidence boundary can be computed via the quantile function of the χ^2-distribution.

A prospective inequality concerning the Kullback–Leibler divergence and geodesic distance was derived in section 4.7. However, a more detailed investigation is required to confirm the validity of this result.

Overall, section 5 consists of applications of the developed methods to practical examples. Initially, they are applied to a toy model in section 5.1 and the implications of non-linear chart transition are discussed in detail. Furthermore, the effects of this transformation on the shape and size of confidence regions, the geometric density and on curvature are examined.

After this, the methods are employed in the analysis of a cosmological model of the distance–redshift relationship for type 1A supernovæ using real data in section 5.2. The exact confidence regions associated with the SCP dataset were constructed. Further, it was revealed that the

geometric density and scalar curvature fluctuate strongly in the vicinity of the chart boundary which in turn can cause instabilities in the process of constructing confidence boundaries via the integral manifold method.

Finally, the performance of the integral method was compared against conventional means of determining the exact confidence boundary in section 5.3 where it was discussed that the implemented integral manifold method typically outperforms naïve grid sampling schemes in terms of accuracy by orders of magnitude. Moreover, due to the respective scaling properties of both schemes, this relative performance advantage of the integral manifold method was observed to become even more pronounced for higher-dimensional parameter spaces.

6.2 Outlook and Future

While this thesis has produced several interesting results, it was of course not possible to address all open questions due to the given time constraints. In the following, I want to summarise some avenues of inquiry which I personally consider to be especially interesting or anticipate to yield fruitful results. These broadly fall into three categories:

1. simple questions which can be answered through straightforward application of the methods and results discussed in this thesis,
2. questions which would require an extension of the presented formalism,
3. speculative ideas which may or may not have a well-defined answer.

Firmly rooted in the first category is the numerical implementation of the integral manifold method for arbitrary likelihoods. Such an implementation would also enable one to study the geometries of more complicated asymmetric error distributions for which an analytical expression of the Kullback–Leibler divergence is not known. Particularly, error distributions given by multivariate t-distributions for have direct applications in experimental sciences.

It should be studied in detail how the general size and volume of confidence regions changes if new parameters are added to an effective model or if uninformative parameters are removed from the model. Likewise, the effect of model reductions on the predictions in form of the confidence bands should be considered in detail.

As pointed out, it would be instructive to investigate the structure constants and the Killing form of the Lie subalgebra of likelihood-annihilating vector fields more closely, given that the structure constants appear to constitute smooth functions in this case instead of constants.

While the quantification of uncertainty in the predictions of a model were investigated in the form of pointwise confidence bands, the construction of simultaneous confidence bands has not been addressed. Further, the construction of prediction bands would require the Bayesian analogue of confidence regions.

Overall, little attention was paid to to different choices of priors and how their use affects information-geometric analyses of datasets in a Bayesian context. An example of a Bayesian analogue of the Fisher information matrix was shown e.g. in [69].

While [30, 31] proposed an extension to the usual geometric formalism which accounts for uncertainties in the observations x_i via a correction factor in the Fisher metric, this idea has yet to be extended for non-normal likelihoods.

Given that models which are obtained from physical theories are often non-linear with respect to the parameters, it would be helpful to be able to systematically construct coordinate transformations which decorrelate the parameters and therefore result in charts which are less distorted and thus exhibit more rounded confidence regions. This might be possible from considerations of either the embedding condition or the gradient of the geometric density. Moreover, it is not known how the boundedness of parameters in form of charts can be accounted for in the definition of non-linear confidence regions for $\dim \mathcal{M} > 1$. Theoretically, it should be possible to find a chart in which the confidence regions take a perfectly ellipsoidal form if the Ricci scalar vanishes on the manifold, i.e. if it is flat.

In [2, 3] it is often stressed that the asymmetry of information divergences is a key property. While Riemannian geometry can only encode the symmetric behaviour of the Kullback–Leibler divergence, the more general Finsler metric can be used to describe geometries where distances are asymmetric. Therefore, a Finsler geometry might be able to provide a more accurate approximation to the Kullback–Leibler divergence and incorporate more of the structural information it contains, which could lead to novel results. This idea is also discussed for example in [72].

It could be possible to reconstruct the likelihood function approximately from known parametrisations of confidence boundaries. Given that they encode significantly more information about the distortion compared with a simple covariance matrix, it is more than plausible that such an approximation might yield more accurate results than assuming that the likelihood is normal and can be modelled using the Fisher metric as evaluated at the MLE.

Appendix

Evaluation of Models on the Confidence Boundary

The topological relation to be proven for a continuous map $f : \mathcal{M} \longrightarrow \mathcal{Z}$ is given by

$$\partial f(C) \subseteq f(\partial C) \tag{6.1}$$

for some set $C \subseteq \mathcal{M}$. First, it is necessary to recall that a map f between topological spaces $(\mathcal{M}, \mathcal{O}_\mathcal{M})$ and $(\mathcal{Z}, \mathcal{O}_\mathcal{Z})$ is said to be closed if it always maps closed sets in the domain to closed sets in the target. This property can alternatively be stated as

$$\overline{f(C)} \subseteq f(\overline{C}). \tag{6.2}$$

Since the assumed continuity of f also implies the opposite direction of this inclusion, the two sides are actually equal in this case.

Next, one proceeds with a proof by contradiction, that is, one assumes that $\partial f(C) \setminus f(\partial C) \neq \varnothing$. Then there must exist a $p \in \mathcal{Z}$ such that

$$\begin{aligned}
& p \in \partial f(C) && \wedge && p \notin f(\partial C) \\
\iff \quad & p \in \overline{f(C)} \setminus \operatorname{Int}\Big(f(C)\Big) && \wedge && p \notin f(\partial C) \\
\iff \quad & p \in \overline{f(C)} && \wedge \quad p \notin \operatorname{Int}\Big(f(C)\Big) && \wedge \quad p \notin f(\partial C)
\end{aligned}$$

where $\operatorname{Int}(A)$ denotes the interior of a set A. Since f is a continuous closed map, $\overline{f(C)} = f(\overline{C})$, therefore

$$\begin{aligned}
& p \in \overline{f(C)} && \wedge && p \notin f(\partial C) && \wedge && p \notin \operatorname{Int}\Big(f(C)\Big) \\
\iff \quad & p \in f(\overline{C}) && \wedge && p \notin f(\partial C) && \wedge && p \notin \operatorname{Int}\Big(f(C)\Big) \\
\iff \quad & && p \in f(\overline{C}) \setminus f(\partial C) && \wedge && p \notin \operatorname{Int}\Big(f(C)\Big).
\end{aligned}$$

It is always true that $f(A) \setminus f(B) \subseteq f(A \setminus B)$ with equality if and only if f is injective. Thus the statement is still valid if the set on the left-hand side is enlarged

$$\begin{aligned}
& p \in f(\overline{C}) \setminus f(\partial C) && \wedge && p \notin \operatorname{Int}\Big(f(C)\Big) \\
\implies \quad & p \in f(\overline{C} \setminus \partial C) && \wedge && p \notin \operatorname{Int}\Big(f(C)\Big) \\
\iff \quad & p \in f(\operatorname{Int} C) && \wedge && p \notin \operatorname{Int}\Big(f(C)\Big)
\end{aligned}$$

where the last statement creates a contradiction if $f(\operatorname{Int} C) \subseteq \operatorname{Int}\Big(f(C)\Big)$ holds, which is precisely the definition of an open map f. Thus, for continuous maps f which are both open and closed, the relation $\partial f(C) \subseteq f(\partial C)$ must hold.

While this is certainly a valid and illuminating result, having to prove topological openness and closedness every time a new model map is studied can be rather tedious. Hence, it would be

advantageous to have a slightly stronger but more practical criterion which model maps can be tested for and from which it already follows that the map is both open and closed.

With this in mind, the target space $\mathcal{Z}$ is now considered in more detail. Once evaluated at a parameter configuration $\theta \in \mathcal{M}$, the model map is a function $y_{\text{model}}(\,\cdot\,;\theta) : \mathcal{X} \longrightarrow \mathcal{Y}$. As argued in section 3.2, one can restrict one's attention to model maps which are continuous with respect to the observation conditions $x \in \mathcal{X}$ without harm. That is, the target space $\mathcal{Z}$ is given by

$$\mathcal{Z} = C^0(\mathcal{X}, \mathcal{Y}) = \left\{ y_{\text{model}}(\,\cdot\,;\theta) : \mathcal{X} \longrightarrow \mathcal{Y} \;\middle|\; y_{\text{model}}(\,\cdot\,;\theta) \text{ continuous} \right\}. \tag{6.3}$$

By the definition of global structural identifiability established in section 3.6, every parameter configuration $\theta \in \mathcal{M}$ must produce a unique prediction $y_{\text{model}}(\,\cdot\,;\theta) : \mathcal{X} \longrightarrow \mathcal{Y}$. As a result, a map $f : \mathcal{M} \longrightarrow C^0(\mathcal{X}, \mathcal{Y})$ which is globally structurally identifiable on a set $C \subseteq \mathcal{M}$ must also be injective on C. In addition, by restricting the target of f to $\mathcal{N} \coloneqq f(C) \subset \mathcal{Z}$, the map trivially becomes surjective onto $\mathcal{N} = f(C)$ such that $f : C \longrightarrow f(C)$ is bijective overall.

It is well-known that bijective maps are open if and only if they are closed: if $h : (\mathcal{M}, \mathcal{O}_{\mathcal{M}}) \longrightarrow (\mathcal{N}, \mathcal{O}_{\mathcal{N}})$ is bijective, one has that

$$\forall U \in \mathcal{O}_{\mathcal{M}} : \qquad h(\underbrace{U^c}_{\text{closed}}) = h(\mathcal{M} \setminus U) \overset{\text{injective}}{=} h(\mathcal{M}) \setminus h(U) \overset{\text{surjective}}{=} \underbrace{\mathcal{N} \setminus \overbrace{h(U)}^{\text{open}}}_{\Rightarrow \text{ closed}} \tag{6.4}$$

where $h(U)$ must be open because $U \in \mathcal{O}_{\mathcal{M}}$ and h is an open map by assumption. Therefore, h must be closed. The opposite direction can be shown by a similar argument.

Further, it is clear that the openness of a bijective (i.e. invertible) map h is equivalent to the requirement that the inverse map h^{-1} is continuous since this means that the preimages of open sets are open. Namely, since h^{-1} exists, one has

$$\forall U \in \mathcal{O}_{\mathcal{M}} : \quad h(U) \in \mathcal{O}_{\mathcal{Z}} \qquad \Longleftrightarrow \qquad \forall U \in \mathcal{O}_{\mathcal{M}} : \quad \text{preim}_{h^{-1}}(U) \in \mathcal{O}_{\mathcal{N}} \tag{6.5}$$

where the right-hand side coincides precisely with the requirement that $h^{-1} : \mathcal{N} \longrightarrow \mathcal{M}$ is continuous.

Finally, the so-called "closed map lemma" states that continuous maps from compact spaces into Hausdorff spaces are closed and proper. Although no restrictions were placed on the set $C \subseteq \mathcal{M}$ up to this point, its compactness is generally given because the metric topology induced on $\mathcal{M}$ is equivalent to the standard topology and confidence regions $\mathcal{C}_q \subseteq \mathcal{M}$ of a level $q < 1$ are typically bounded. Therefore, if one can establish a topology on $f(C) \subset C^0(\mathcal{X}, \mathcal{Y})$ which is Hausdorff and renders f continuous, the closed map lemma immediately implies that f is open, closed and a homeomorphism.

The finest topology which renders $f : C \subseteq \mathcal{M} \longrightarrow f(C) \subset C^0(\mathcal{X}, \mathcal{Y})$ continuous is the final topology $\mathcal{O}_{\text{final}}$ which in this case is given by

$$\mathcal{O}_{\text{final}} = \left\{ V \subseteq f(C) \;\middle|\; \text{preim}_f(V) \in \mathcal{O}_{\mathcal{M}} \right\}. \tag{6.6}$$

That is to say, if even a single element were added to $\mathcal{O}_{\text{final}}$, then f would no longer be continuous. In this particular case, where f is bijective, it is straightforward to see that $\left(f(C), \mathcal{O}_{\text{final}}\right)$ is Hausdorff if $(C, \mathcal{O}_C)$ is Hausdorff. That is, due to the injectivity of f, one has

$$\forall p, q \in f(C): \qquad p \neq q \qquad \Longrightarrow \qquad f^{-1}(p) \neq f^{-1}(q). \tag{6.7}$$

By definition, $(C, \mathcal{O}_C)$ being Hausdorff implies

$$\exists U_p, U_q \in \mathcal{O}_C : f^{-1}(p) \in U_p : f^{-1}(q) \in U_q : \qquad U_p \cap U_q = \varnothing \tag{6.8}$$

which yields again by injectivity and by choice of the final topology that

$$U_p \cap U_q = \varnothing \qquad \Longrightarrow \qquad f(U_p) \cap f(U_q) = \varnothing \quad \wedge \quad f(U_p), f(U_q) \in \mathcal{O}_{\text{final}}. \tag{6.9}$$

Since the topology $\mathcal{O}_{\mathcal{M}}$ is induced by the open balls of the geodesic distance via the Fisher metric and the subset topology $\mathcal{O}_{\mathcal{M}}|_C$ preserves the Hausdorff property, the final topology $\mathcal{O}_{\text{final}}$ is therefore also Hausdorff.

This concludes the proof that the compactness of $C \subseteq \mathcal{M}$ and globally structurally identifiability (i.e. injectivity) of models $f : C \subseteq \mathcal{M} \longrightarrow f(C) \subset C^0(\mathcal{X}, \mathcal{Y})$ guarantee that said models are continuous open and closed, wherefore $\partial f(C) \subseteq f(\partial C)$ holds.

Furthermore, since bijective maps which are continuous and open must be homeomorphisms, one recognises that the restricted target $f(C)$ constitutes a $(\dim \mathcal{M})$-dimensional topological subspace embedded in $C^0(\mathcal{X}, \mathcal{Y})$. Curiously, whereas continuity is unaffected if the topology on the target space is made coarser (i.e. if sets are removed), the reverse is true for openness—the openness of a map is preserved if the topology on the target space is made finer (i.e. if more sets are made open in the target). This delicate balance suggests that there is not a great deal of freedom in the choice of a topology on $f(C) \subset C^0(\mathcal{X}, \mathcal{Y})$ such that f is both continuous as well as open (and subsequently closed due to bijectivity).

Numerical Computation of Derivatives

The derivative of a differentiable[72] scalar function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is analytically defined as

$$f'(x) := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}. \tag{6.10}$$

Traditionally, the numerical computation of derivatives is performed by approximating the limit $h \to 0$ using values on the order of $h \approx 10^{-8}$ in the definition which is then referred to as the forward difference quotient.

However, this finite difference approach has several drawbacks: of course, if the parameter h is too large then the forward difference quotient provides a poor approximation to the true derivative. On

[72] Although it will not be explicitly stated throughout, it must certainly be assumed that all functions considered in the context of this section are suitably differentiable or holomorphic wherever necessary.

the other hand, there is a lower bound on the value of h posed by the precision of the floating-point arithmetic that is employed in numerical calculations and the rounding errors that come with it. This quantity is also referred to as "machine epsilon" and takes the value of $2^{-52} \approx 2.2 \cdot 10^{-16}$ for signed 64-bit floating-point numbers.[73]

Clearly, this loss of precision is amplified even more when computing higher order derivatives of a function. Although this is an inherent problem to the finite difference method and ultimately cannot be fixed, there are several ways to combat this loss of precision and make the finite difference scheme more robust against round-off errors.

The simplest adaptation of the forward difference quotient is the so-called centered difference quotient

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x-h)}{2h}. \tag{6.11}$$

In a sense, the forward difference quotient actually approximates $f'(x + h/2)$ instead of $f'(x)$ for finite h. Thus, the centered difference quotient improves the approximation to the derivative at the actual position of interest for finite h by its symmetry, at no additional computational cost.

One can further improve the accuracy of a numerical derivative by evaluating the function at more points and interpolating the desired value by forming a linear combination with appropriate weights[74]

$$f'(x) \approx \frac{-f(x+2h) + 8f(x+h) - 8f(x-h) + f(x-2h)}{12h}. \tag{6.12}$$

While there is in principle no upper limit on the number of additional function evaluations one can use to stabilise the difference quotient, one inevitably hits a point of diminishing returns where the gained accuracy is not worth the extra computational effort.

Lastly, there is also the imaginary step method

$$f'(x) = \lim_{h \to 0} \frac{\operatorname{Im}\{f(x + \mathrm{i}h)\}}{h} \tag{6.13}$$

where the analytical continuation of the original function is evaluated on the complex plane perpendicular to the desired point on the real line. As it turns out, this method yields the best results out of all the finite difference schemes presented here (see the comparison in figure 57) as no additions or subtractions between function values are necessary.

Apart from the fact that one is restricted to real-valued functions in using this method, consideration

[73] Specifically, this is the distance between the 64-bit floating-point value representation of 1.0 and the next larger representable value in Julia as per the documentation.

[74] To aid in finding the correct coefficients this linear combination, there exist convenient tools such as [75] provided by C. Taylor.

of the Taylor expansion of $f(x + \mathrm{i}h)$ for small $\mathrm{i}h$ reveals

$$f(x+\mathrm{i}h) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (\mathrm{i}h)^n = f(a) + \mathrm{i}hf'(a) + \frac{\mathrm{i}^2 h^2}{2!} f''(a) + \frac{\mathrm{i}^3 h^3}{3!} f'''(a) + \ldots \tag{6.14}$$

$$= f(a) + \mathrm{i}hf'(a) - \frac{h^2}{2!} f''(a) - \frac{\mathrm{i}h^3}{3!} f'''(a) + \frac{h^4}{24} f''''(a) + \ldots \tag{6.15}$$

$$= \left[f(a) - \frac{h^2}{2!} f''(a) + \ldots \right] + \mathrm{i} \left[hf'(a) - \frac{h^3}{3!} f'''(a) + \ldots \right]. \tag{6.16}$$

Therefore, although it outperforms other finite difference methods by a large margin, the imaginary step method still introduces a slight but noticeable rounding error in the final division step. However, it likely served as the inspiration for the use of dual numbers in numerical differentiation.

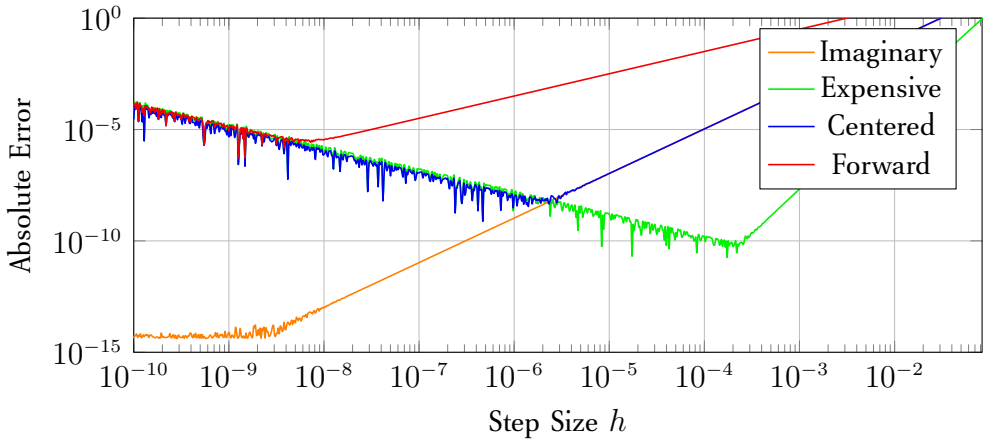


Figure 57: Comparison of the accuracies achieved by different numerical differentiation schemes as a function of the finite differencing parameter h. It reveals that on the one hand, large h provide a poor approximation to the derivative while on the other hand very small step sizes introduce rounding errors due to the finite precision. As a result, basically all finite differencing schemes feature an optimal value of h at which their error is minimal.

As a last resort, one can of course circumvent the rounding errors caused in finite difference schemes by using arbitrary precision floating-point arithmetic at the expense of increased requirements on memory and computational time.

Dual Numbers and Automatic Differentiation

Similar to how the real number field $\mathbb{R}$ can be extended to the complex numbers $\mathbb{C} = \{a + \mathrm{i}b \mid a, b \in \mathbb{R}\}$ using the imaginary unit i which satisfies

$$\mathrm{i}^2 = -1 \qquad \text{but} \qquad \mathrm{i} \neq \sqrt{-1}, \tag{6.17}$$

one can also extend the real numbers to the dual numbers $\mathbb{D} = \{a + \epsilon b \mid a, b \in \mathbb{R}\}$ using the so-called dual unit ϵ with the properties

$$\epsilon^2 = 0 \qquad \text{but} \qquad \epsilon \neq 0. \tag{6.18}$$

With this property, the arithmetic rules for dual numbers work out to

$$(a + \epsilon b) +_{\mathbb{D}} (c + \epsilon d) = (a + c) + \epsilon(b + d) \tag{6.19}$$

$$(a + \epsilon b) \cdot_{\mathbb{D}} (c + \epsilon d) = ac + \epsilon(ad + bc) + \underbrace{bd\epsilon^2}_{=0}. \tag{6.20}$$

One can convince oneself that this algebraic structure does indeed have the desired effect of "automatically" carrying the derivative of a function throughout a calculation using elementary examples such as

$$(a + \epsilon)^4 = a^4 + 4a^3\epsilon + \underbrace{6a^2\epsilon^2 + 4a\epsilon^3 + \epsilon^4}_{=0}. \tag{6.21}$$

Analytically extending the original function f to the dual numbers, evaluating at $(a + \epsilon b) \in \mathbb{D}$ and expanding it as a Taylor series in terms of ϵb yields

$$f(a + \epsilon b) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (\epsilon b)^n = f(a) + \epsilon b\, f'(a) + \underbrace{\frac{\epsilon^2 b^2}{2} f''(a)}_{=0} + \underbrace{\frac{\epsilon^3 b^3}{6} f'''(a) + \ldots}_{=0} \tag{6.22}$$

$$= f(a) + \epsilon b\, f'(a). \tag{6.23}$$

Thus, in one evaluation of the function, both its value and the value of its derivative are computed. Specifically, from this one defines the automatic derivative as

$$f'(x) = \text{Dual}\Big\{ f(x + \epsilon) \Big\} \tag{6.24}$$

where no approximation in the form of choosing a finite difference parameter needs to be made. Thus, the automatic derivative is generally accurate to machine precision.

Although dual numbers yield a practical numerical implementation of differentiation, they do not play a meaningful role in analytical mathematics due to the fact that, in contrast to the complex numbers, they only constitute a commutative ring, but not a number field. Namely, because of the defining property $\epsilon^2 = 0$, the dual number ring lacks multiplicative inverses to elements of the form $0 + b\epsilon$ for all $b \in \mathbb{R}$.

For higher order derivatives, one can extend this system further to hyper-dual numbers and use them to compute derivatives of functions to an arbitrary degree. Alternatively, higher derivatives can be computed recursively by allowing the coefficients of a and b in a dual number $(a + \epsilon b)$ to also be dual numbers.

Taking advantage of Julia's optionally typed system, an implementation of this automatic differentiation scheme is already provided by the `ForwardDiff.jl` package, which allows for the convenient

automatic computation of derivatives, gradients, Jacobians and Hessians. Specifically, to differentiate maps whose domain is some subset of $\mathbb{R}^n$, the analogous extension to $\mathbb{D}^n$ is used, where the map is then evaluated.

Technical Details of Implementation

Instead of providing pseudocode for algorithms in this section, examples will be given in plain Julia code. The reasons for this choice are twofold: due to its high-level character, Julia code already rivals pseudocode in its terseness and simplicity. On top of that, it has the added benefit of being directly executable and thus it avoids any kind of ambiguity that is sometimes present in pseudocode. Accordingly, it can be argued that this is well worth sacrificing true language agnosticism.

Although the depicted algorithms are executable without further modification, they represent significantly simplified version where any type restrictions or optimisation measures have been omitted for the sake of brevity. Certainly, not only the Julia compiler but also the architectures of any packages involved may be subject to change in the future.

Immediately following the publication of this thesis, an implementation of the methods that were developed and used for analyses will be published and maintained for the foreseeable future as an open source package for the Julia language. Most likely, it will be registered under the name `InformationGeometry.jl`. As a result, the ability to compute exact confidence regions easily and efficiently will be available for free to anyone in the broader research community instead of being restricted to a small group of experts in the field of information geometry.

Concrete Layout of the Implemented Programme

In its current implementation, the specialised data type `DataSet` holds the $(x_{\text{data}}, y_{\text{data}}, \Sigma)$ triple while the data type `DataModel` also carries the model function and its first derivatives on top of that. These containers can then be passed between specialised functions in a convenient fashion to compute likelihoods, embeddings, the Fisher metric and any other secondary quantities like geodesics, boundaries of confidence regions and so on. A simplified example of how this is achieved is shown in algorithm 1.

As finding the boundary of a confidence region mainly consists of solving a system of ordinary differential equations to find the integral curves or surfaces of a vector field, it is usually advisable to use pre-existing packages for this task in order to obtain the best results. A significant amount of progress has been made in the development of efficient algorithms for solving differential equations over the last 40 years. Besides sophisticated interpolation methods, many modern solver algorithms can also be rigorously proven to satisfy certain stability conditions.

Apart from offering a large range of different numerical solvers for all kinds of different situations, the `DifferentialEquations.jl` package has a handful of advanced features which improve overall accuracy and execution time. For instance, the so-called “callback” feature can be used

```
struct DataSet
    x::AbstractVector
    y::AbstractVector
    sigma::AbstractMatrix
end
```

```
struct DataModel
    Data::DataSet
    model::Function
    dmodel::Function
end
```

```
function loglikelihood(DM::DataModel, params::Vector{<:Real})
    resid(i) = (DM.Data.y[i] .- DM.model(DM.Data.x[i],params))/DM.Data.sigma[i])
    R = sum( dot(resid(i),resid(i)) for i in 1:length(DM.Data.x) )
    return -0.5*(length(DM.Data.x)*log(2pi) + 2*sum(log.(DM.Data.sigma)) + R)
end
```

Algorithm 1: Using suitable constructors, the consistency of the arguments can be verified.

to check whether a condition is fulfilled during the process of numerically solving a differential equation and subsequently to terminate or alter this solution process in a controlled fashion. In the context of determining confidence boundaries, one can use this to terminate the integration process once an integral curve reaches its starting point again. Most importantly, this can be used to constrain particles, i.e. the solution, to certain volumes or surfaces in position space or phase space. Furthermore, one can enforce quantities to remain conserved during the solution process using the manifold projection feature which improves the stability of the solving process. In a physical context, this could for example be used to ensure conservation of momentum and energy or more exotic dynamical symmetries of a system such as conservation of the Runge–Lenz vector. By using the value of the log-likelihood as a conservation quantity, one can ensure that the integral manifold representing the confidence boundary really is an iso-likelihood surface.

A simplified implementation of the integral manifold scheme using the termination condition as well as the manifold projection in the generation of the confidence boundary for a two-dimensional parameter manifold is depicted in in algorithm 2.

```
function GenerateBoundary(DM::DataModel, u0::Vector; tol=1e-14)
    LogLikeOnBoundary = loglikelihood(DM,u0)
    function IntegralCurveODE(du,u,p,t)
        du .= OrthVF(DM,u)
    end
    function IsoLikelihood(resid,u,p,t)
        resid[1] = loglikelihood(DM,u) - LogLikeOnBoundary
    end
    TerminateCondition(u,t,integrator) = u[2] - u0[2]
    cb = CallbackSet(ManifoldProjection(IsoLikelihood),
    ContinuousCallback(TerminateCondition,terminate!,nothing))
    tspan = (0.0,1000.0)
    prob = ODEProblem(IntegralCurveODE,u0,tspan)
    return solve(prob,Tsit5(),reltol=tol,abstol=tol,callback=cb)
end
```

Algorithm 2: Example of an implementation for the generation of confidence boundaries in the form of integral curves given a starting point `u0`. For models depending on three or more parameters, the condition for termination must be adapted. Also, since it is typically desirable to mesh higher-dimensional confidence boundaries using integral curves which lie in planar slices of the confidence region, the likelihood-annihilating vector field denoted by `OrthVF()` must be projected appropriately.

References

[1] Supernova Cosmology Project (Union2.1) Dataset. http://supernova.lbl.gov/Union/figures/SCPUnion2.1_mu_vs_z.txt

[2] Amari, S. : *Information Geometry and Its Applications*. Springer Japan (Applied Mathematical Sciences). https://books.google.de/books?id=UkSFCwAAQBAJ. – ISBN 9784431559788

[3] Amari, S. ; Nagaoka, H. : *Methods of Information Geometry*. American Mathematical Society (Translations of mathematical monographs). https://books.google.de/books?id=vc2FWSo7wLUC. – ISBN 9780821843024

[4] Amari, S. ; Tsuchiya, N. ; Oizumi, M. : Geometry of Information Integration. (2017). https://arxiv.org/pdf/1709.02050v1.pdf

[5] Amari, S.-I. ; Barndorff-Nielsen, O. E. ; Kass, R. E. ; Lauritzen, S. L. ; Rao, C. R. ; Gupta, S. S. (Hrsg.): *Lecture Notes–Monograph Series*. Bd. Volume 10: *Chapter 5: Differential Metrics in Probability Spaces*. Institute of Mathematical Statistics. – 217–240 S. http://dx.doi.org/10.1214/lnms/1215467062. http://dx.doi.org/10.1214/lnms/1215467062

[6] Anastasiou, A. ; Ley, C. : Bounds for the asymptotic normality of the maximum likelihood estimator using the Delta method. (2015). https://arxiv.org/pdf/1508.04948v3.pdf

[7] Anastasiou, A. ; Reinert, G. : Bounds for the normal approximation of the maximum likelihood estimator. (2014). https://arxiv.org/pdf/1411.2391v3.pdf

[8] Athreya, K. ; Lahiri, S. : *Measure Theory and Probability Theory*. Springer (Springer Texts in Statistics). https://books.google.de/books?id=9tv0taI8l6YC. – ISBN 9780387329031

[9] Ballmann, W. : Critical Point Theory of the Energy Functional on Path Spaces. (2015). http://people.mpim-bonn.mpg.de/hwbllmnn/archiv/energyfun1501.pdf

[10] Bartelmann, M. : *Das kosmologische Standardmodell: Grundlagen, Beobachtungen und Grenzen*. Springer Berlin Heidelberg https://books.google.de/books?id=o4CvDwAAQBAJ. – ISBN 9783662596272

[11] Bauckhage, C. : Computing the Kullback-Leibler Divergence between two Generalized Gamma Distributions. In: *CoRR* abs/1401.6853 (2014). http://arxiv.org/abs/1401.6853

[12] Beem, J. ; Ehrlich, P. ; Easley, K. : *Global Lorentzian Geometry, Second Edition*. Taylor & Francis (Chapman & Hall/CRC Pure and Applied Mathematics). https://books.google.de/books?id=N9v-F0VRQR0C. – ISBN 9780824793241

[13] Bender, C. ; Orszag, S. : *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*. Springer (Advanced Mathematical Methods for Scientists and Engineers). https://books.google.de/books?id=-yQXwhE6iWMC. – ISBN 9780387989310

[14] Berger, J. : *Statistical Decision Theory and Bayesian Analysis*. Springer New York (Springer Series in Statistics). https://books.google.de/books?id=1CDaBwAAQBAJ. – ISBN 9781475742862

[15] Besançon, M. ; Anthoff, D. ; Arslan, A. ; Byrne, S. ; Lin, D. ; Papamarkou, T. ; Pearson, J. : Distributions.jl: Definition and Modeling of Probability Distributions in the JuliaStats Ecosystem. In: *arXiv e-prints* (2019), Jul, S. arXiv:1907.08611

[16] Campbell, L. L.: An extended Čencov characterization of the information metric. In: *Proc. Amer. Math. Soc.* 98 (1986), Nr. 1, 135–141. http://dx.doi.org/10.2307/2045782. – DOI 10.2307/2045782. – ISSN 0002–9939

[17] Casella, G. ; Berger, R. : *Statistical Inference*. Brooks/Cole Publishing Company (Duxbury advanced series). https://books.google.de/books?id=nA_vAAAAMAAJ. – ISBN 9780534119584

[18] Caticha, A. : The Basics of Information Geometry. (2014). https://arxiv.org/pdf/1412.5633v1.pdf

[19] Chyzak, F. ; Nielsen, F. : A closed-form formula for the Kullback-Leibler divergence between Cauchy distributions. In: *CoRR* abs/1905.10965 (2019). http://arxiv.org/abs/1905.10965

[20] Crane, K. : Discrete Differential Geometry: An Applied Introduction. (2020). http://www.cs.cmu.edu/~kmcrane/Projects/DDG/paper.pdf

[21] De Maesschalck, R. ; Jouan-Rimbaud, D. ; Massart, D. : The Mahalanobis distance. In: *Chemometrics and Intelligent Laboratory Systems* 50 (2000), Nr. 1, 1 - 18. http://dx.doi.org/https://doi.org/10.1016/S0169-7439(99)00047-7. – DOI https://doi.org/10.1016/S0169-7439(99)00047-7. – ISSN 0169–7439

[22] Engelking, R. : *General Topology*. Heldermann Verlag Berlin, 1989. – ISBN 3–88538–006–4

[23] Fisher, R. A. ; Russell, E. J.: On the mathematical foundations of theoretical statistics. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222 (1922), Nr. 594-604, 309-368. http://dx.doi.org/10.1098/rsta.1922.0009. – DOI 10.1098/rsta.1922.0009

[24] Giesel, E. S.: Investigation of Non-Gaussian Likelihoods in the Framework of Information Geometry.

[25] Gänsicke, B. T. ; Koester, D. ; Raddi, R. ; Toloza, O. ; Kepler, S. O.: SDSSJ124043.01+671034.68: the partially burned remnant of a low-mass white dwarf that underwent thermonuclear ignition? In: *Monthly Notices of the Royal Astronomical Society* 496 (2020), 06, Nr. 4, 4079-4086. http://dx.doi.org/10.1093/mnras/staa1761. – DOI 10.1093/mnras/staa1761. – ISSN 0035–8711

[26] Gray, R. : *Entropy and Information Theory*. Springer New York https://books.google.de/books?id=ZoTSBwAAQBAJ. – ISBN 9781475739824

[27] Guerrero-Cusumano, J.-L. : An asymptotic test of independence for multivariate t and Cauchy random variables with applications. In: *Information Sciences* 92 (1996), Nr. 1, 33 - 45. http://dx.doi.org/https://doi.org/10.1016/0020-0255(96)00036-9. – DOI https://doi.org/10.1016/0020–0255(96)00036–9. – ISSN 0020–0255

[28] Guerrero-Cusumano, J.-L. : A measure of total variability for the multivariate t distribution with applications to finance. In: *Information Sciences* 92 (1996), Nr. 1, 47 - 63. http://dx.doi.org/https://doi.org/10.1016/0020-0255(96)00044-8. – DOI https://doi.org/10.1016/0020–0255(96)00044–8. – ISSN 0020–0255

[29] Hall, B. C.: *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*. Springer (Graduate Texts in Mathematics). https://books.google.de/books?id=m1VQi8HmEwcC. – ISBN 9780387401225

[30] 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

[31] 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

[32] Hoffman, K. ; Kunze, R. : *Linear Algebra (2nd Edition)*. Prentice-Hall Of India Pvt. Limited https://books.google.de/books?id=SeBIPgAACAAJ. – ISBN 9788120302709

[33] Isham, C. J.: *Modern Differential Geometry for Physicists*. Allied Publ. (World Scientific lecture notes in physics). https://books.google.de/books?id=DCn9bjBe27oC. – ISBN 9788177643169

[34] Jaynes, E. ; Bretthorst, G. : *Probability Theory: The Logic of Science*. Cambridge University Press https://books.google.de/books?id=UjsgAwAAQBAJ. – ISBN 9781139435161

[35] Jech, T. : *Set Theory*. Springer, 2007. – ISBN 9783540447610

[36] Jutho ; ho-oto ; getzdan ; Lyon, S. ; Morley, A. ; Privett, A. ; Iouchtchenko, D. ; Saba, E. ; Otto, F. ; Garrison, J. ; TagBot, J. ; Leo ; S., M. P. ; Hauru, M. : Jutho/TensorOperations.jl: v3.0.0. (2020), June. http://dx.doi.org/10.5281/zenodo.3874973. – DOI 10.5281/zenodo.3874973

[37] Kapfer, S. : Computerphysik und Numerische Methoden. (2016)

[38] Keener, R. : *Theoretical Statistics: Topics for a Core Course*. Springer New York (Springer Texts in Statistics). https://books.google.de/books?id=aVJmcega44cC. – ISBN 9780387938394

[39] Kendall, M. ; Stuart, A. : *The Advanced Theory of Statistics*. C. Griffin (The Advanced Theory of Statistics Bd. 1). https://books.google.de/books?id=LxfvAAAAMAAJ

[40] Kullback, S. : *Information Theory and Statistics*. Wiley (A Wiley publication in mathematical statistics). https://books.google.de/books?id=c5lFzQEACAAJ

[41] Lebanon, G. : An Extended Čencov-Campbell Characterization of Conditional Information Geometry. (2012). https://arxiv.org/ftp/arxiv/papers/1207/1207.4139.pdf

[42] Lee, J. M.: *Introduction to Smooth Manifolds*. Springer, 2012. http://dx.doi.org/10.1007/978-1-4419-9982-5. http://dx.doi.org/10.1007/978-1-4419-9982-5

[43] Liu, W. ; Lin, S. ; Piegorsch, W. W.: Construction of exact simultaneous confidence bands for a simple linear regression model. (2008), März. http://dx.doi.org/10.1111/j.1751-5823.2007.00027.x. – DOI 10.1111/j.1751-5823.2007.00027.x

[44] Machta, B. B. ; Chachra, R. ; Transtrum, M. K. ; Sethna, J. P.: Parameter Space Compression Underlies Emergent Theories and Predictive Models. (2013). https://arxiv.org/pdf/1303.6738v1.pdf

[45] Metzger, W. : *Statistical Methods in Data Analysis*. Fakulteit der Natuurwetenschappen, Katholieke Universiteit Nijmegen https://books.google.de/books?id=Tdv4ZwEACAAJ

[46] Meusburger, C. ; Neeb, K.-H. : Topologie. (2019)

[47] Misner, C. ; Thorne, K. ; Wheeler, J. ; Kaiser, D. : *Gravitation*. Princeton University Press https://books.google.de/books?id=zAAuDwAAQBAJ. – ISBN 9781400889099

[48] Mogensen, P. K. ; Riseth, A. N.: Optim: A mathematical optimization package for Julia. In: *Journal of Open Source Software* 3 (2018), Nr. 24, S. 615. http://dx.doi.org/10.21105/joss.00615. – DOI 10.21105/joss.00615

[49] Murray, M. ; Rice, J. : *Differential Geometry and Statistics*. Taylor & Francis (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). https://books.google.de/books?id=ZBa7F9LrDrMC. – ISBN 9780412398605

[50] Nadarajah, S. ; Kotz, S. : Mathematical Properties of the Multivariate t Distribution. In: *Acta Applicandae Mathematica* (2005). http://dx.doi.org/10.1007/s10440-005-9003-4. – DOI 10.1007/s10440-005-9003-4

[51] 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

[52] Nielsen, F. : An elementary introduction to information geometry. (2018). https://arxiv.org/pdf/1808.08271.pdf

[53] Nielsen, F. : On the Kullback-Leibler divergence between location-scale densities. (2019). https://arxiv.org/pdf/1904.10428v2.pdf

[54] Nowaczyk, N. : Geodesics, Energy and Variations. (2009). https://nikno.de/wp-content/uploads/2016/07/geonrgvar.pdf

[55] Onzon, E. : Multivariate Cramér–Rao inequality for prediction and efficient predictors. In: *Statistics & Probability Letters* 81 (2011), Nr. 3, 429 - 437. http://dx.doi.org/https://doi.org/10.1016/j.spl.2010.12.007. – DOI https://doi.org/10.1016/j.spl.2010.12.007. – ISSN 0167–7152

[56] Pearson, K. : On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50 (1900), Nr. 302, 157-175. http://dx.doi.org/10.1080/14786440009463897. – DOI 10.1080/14786440009463897

[57] Penrose, R. : *The Road to Reality: A Complete Guide to the Laws of the Universe*. Vintage Books, 2007. – ISBN 9780679776314

[58] Rackauckas, C. ; Nie, Q. : DifferentialEquations.jl – A Performant and Feature-Rich Ecosystem for Solving Differential Equations in Julia. In: *Journal of Open Source Software* (2017). http://dx.doi.org/http://doi.org/10.5334/jors.151. – DOI http://doi.org/10.5334/jors.151

[59] Revels, J. ; Lubin, M. ; Papamarkou, T. : Forward-Mode Automatic Differentiation in Julia. In: *arXiv:1607.07892 [cs.MS]* (2016). https://arxiv.org/abs/1607.07892

[60] Rényi, A. : On Measures of Entropy and Information. (1961), 547–561. https://projecteuclid.org/euclid.bsmsp/1200512181

[61] Roth, W. : Inner products on ordered cones. In: *New Zealand Journal of Mathematics* 30 (2001), S. 157–175

[62] Rothenberg, T. J.: Identification in Parametric Models. In: *Econometrica* 39 (1971), Nr. 3, 577–591. http://www.jstor.org/stable/1913267. – ISSN 00129682, 14680262

[63] Schäfer, B. M. ; Douspis, M. ; Aghanim, N. : Implications of bias evolution on measurements of the integrated Sachs-Wolfe effect: errors and biases in parameter estimation. (2009). http://dx.doi.org/10.1111/j.1365-2966.2009.14991.x. – DOI 10.1111/j.1365-2966.2009.14991.x

[64] Schnörr, C. : Information Geometry and Machine Learning. (2018). https://ipa.math.uni-heidelberg.de/dokuwiki/doku.php?id=teaching:st18:infogeo:start

[65] Schuller, F. P.: All spacetimes beyond Einstein (Obergurgl Lectures). (2011). http://arxiv.org/abs/1111.4824v1.pdf

[66] Schuller, F. P.: Heraeus International Winter School on Gravity and Light. (2015). https://www.youtube.com/playlist?list=PLFeEvEPtX_0S6vxxiiNPrJbLu9aK1UVC_

[67] Schuller, F. P.: Lectures on the Geometric Anatomy of Theoretical Physics. (2015). https://www.youtube.com/playlist?list=PLPH7f_7ZlzxTi6kS4vCmv4ZKm9u8g5yic

[68] Seber, G. ; Wild, C. : *Nonlinear Regression*. Wiley (Wiley Series in Probability and Statistics). https://books.google.de/books?id=YBYlCpBNo_cC. – ISBN 9780471471356

[69] Sellentin, E. ; Quartin, M. ; Amendola, L. : Breaking the spell of Gaussianity: forecasting with higher order Fisher matrices. (2014). http://dx.doi.org/10.1093/mnras/stu689. – DOI 10.1093/mnras/stu689

[70] Sellentin, E. ; Schäfer, B. M.: Non-Gaussian forecasts of weak lensing with and without priors. (2015). http://dx.doi.org/10.1093/mnras/stv2805. – DOI 10.1093/mnras/stv2805

[71] Shannon, C. E.: A Mathematical Theory of Communication. In: *Bell System Technical Journal* 27 (1948), Nr. 3, 379-423. http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x. – DOI 10.1002/j.1538–7305.1948.tb01338.x

[72] Shen, Z. : Riemann-Finsler Geometry with Applications to Information Geometry. In: *Chinese Annals of Mathematics, Series B* 27 (2006), 08, S. 73–94. http://dx.doi.org/10.1007/s11401-005-0333-3. – DOI 10.1007/s11401–005–0333–3

[73] Sloane, N. J. A. ; Plouffe, S. : The On-Line Encyclopedia of Integer Sequences. (2020). http://oeis.org

[74] Suzuki, N. ; Rubin, D. ; Lidman, C. ; Aldering, G. ; Amanullah, R. ; Barbary, K. ; Barrientos, L. F. ; Botyanszki, J. ; Brodwin, M. ; Connolly, N. ; Dawson, K. S. ; Dey, A. ; Doi, M. ; Donahue, M. ; Deustua, S. ; Eisenhardt, P. ; Ellingson, E. ; Faccioli, L. ; Fadeyev, V. ; Fakhouri, H. K. ; Fruchter, A. S. ; Gilbank, D. G. ; Gladders, M. D. ; Goldhaber, G. ; Gonzalez, A. H. ; Goobar, A. ; Gude, A. ; Hattori, T. ; Hoekstra, H. ; Hsiao, E. ; Huang, X. ; Ihara, Y. ; Jee, M. J. ; Johnston, D. ; Kashikawa, N. ; Koester, B. ; Konishi, K. ; Kowalski, M. ; Linder, E. V. ; Lubin, L. ; Melbourne, J. ; Meyers, J. ; Morokuma, T. ; Munshi, F. ; Mullis, C. ; Oda, T. ; Panagia, N. ; Perlmutter, S. ; Postman, M. ; Pritchard, T. ; Rhodes, J. ; Ripoche, P. ; Rosati, P. ; Schlegel, D. J. ; Spadafora, A. ; Stanford, S. A. ; Stanishev, V. ; Stern, D. ; Strovink, M. ; Takanashi, N. ; Tokita, K. ; Wagner, M. ; Wang, L. ; Yasuda, N. ; Yee, H. K. C.: The Hubble Space Telescope Cluster Supernova Survey: V. Improving the Dark Energy Constraints Above $z > 1$ and Building an Early-Type-Hosted Supernova Sample. In: *The Astrophysical Journal* 746 (2012), jan, Nr. 1, 85. http://dx.doi.org/10.1088/0004-637x/746/1/85. – DOI 10.1088/0004–637x/746/1/85

[75] Taylor, C. : Finite Difference Coefficients Calculator. http://web.media.mit.edu/~crtaylor/calculator.html

[76] Transtrum, M. K.: Manifold boundaries give "gray-box" approximations of complex models. (2016). https://arxiv.org/pdf/1605.08705v1.pdf

[77] 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

[78] 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

[79] Čencov, N. N.: *Statistical Decision Rules and Optimal Inference*. American Mathematical Society (Translations of mathematical monographs). https://books.google.de/books?id=63CPCwAAQBAJ. – ISBN 9780821813478

[80] Vugrin, K. W. ; Swiler, L. P. ; Roberts, R. M. ; Stucky-Mack, N. J. ; Sullivan, S. P.: Confidence region estimation techniques for nonlinear regression in groundwater flow: Three case studies. In: *Water Resources Research* 43 (2007), Nr. 3. http://dx.doi.org/10.1029/2005WR004804. – DOI 10.1029/2005WR004804

[81] Vugrin, K. W. ; Swiler, L. P. ; Roberts, R. M. ; Stucky-Mack, N. J. ; Sullivan, S. P.: Confidence region estimation techniques for nonlinear regression in groundwater flow: Three case studies. In: *Water Resources Research* 43 (2007), Nr. 3. http://dx.doi.org/10.1029/2005WR004804. – DOI 10.1029/2005WR004804

[82] Wald, R. M.: *General Relativity*. University of Chicago Press, 1984. – ISBN 9780226870328

[83] 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

[84] Wilks, S. S.: The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. In: *Ann. Math. Statist.* 9 (1938), 03, Nr. 1, 60–62. http://dx.doi.org/10.1214/aoms/1177732360. – DOI 10.1214/aoms/1177732360

[85] Wu, H. ; Neale, M. C.: Adjusted confidence intervals for a bounded parameter. In: *Behavior genetics* (2012). http://dx.doi.org/10.1007/s10519-012-9560-z. – DOI 10.1007/s10519-012-9560-z

Acknowledgements

First and foremost, I want to express my gratitude to Professor Björn Malte Schäfer for inviting me into his research group and for supervising this thesis. I very much enjoyed the relaxed work environment and found the atmosphere in the group to be excellent. He showed me by example what it means to go far above and beyond what are generally considered a researcher's obligations to the scientific community. Furthermore, I want to thank him for providing me with the opportunity to continue our research as a PhD student in his group.

Second, I want to thank Professor Klaus Mecke for kindly agreeing to co-supervise and taking the time to referee this thesis.

Also, I am grateful to Frederic P. Schuller who not only first introduced me to the subject of differential geometry several years ago but who also acquainted me with Björn and his research group. His brilliant lecturing instilled in me a deep sense of appreciation for the necessity and benefit of mathematical rigour in theoretical physics and science in general.

Moreover, I want to thank everyone both inside and outside the Schäfer group with whom I have enjoyed fruitful and stimulating discussions and who provided me with insightful comments regarding this thesis. Notably, this in includes—in no particular order—Eileen Giesel, Maximilian Düll, Marie Teich, Leonard Küppers, Brian Kantor, Max Ellinger, Ricardo Waibel and Alena Brändle.

Finally, I want to thank my family for their moral support throughout the years and for enabling me to pursue my passion for science. I am particularly indebted to my mother whose hard work in raising my two younger sisters and myself I greatly appreciate and to whom I dedicate this thesis.

The presented work was carried out during the winter semester 2019 / 2020 and summer semester 2020 at the Center of Astronomy in Heidelberg (ZAH) under the joint supervision of Prof. Klaus Mecke and Prof. Björn Malte Schäfer.

I hereby declare that I have written this Master's thesis on my own and using only the quoted references.

_____

Rafael Arutjunjan