PhD Degree Dissertation

Model Selection for Data Analysis Based on the MDL Principle

Bono Stoychev Nonchev

April, 2015

Sofia University "St. Kliment Ohridski"
Faculty of Mathematics and Informatics
Department of Probability, Operational Research and Statistics

Advisor
Docent Plamen Mateev, Ph.D.

Statistics is an information science, the first and most fully developed information science.

– Bradley Efron

Parts of this dissertation are based on the following papers.

The literature review in Chapter 2 is based on:

> *Recent Advances in MDL Principle*, B. Nonchev.
> Manuscript in preparation.

Chapter 3 and Chapter 4 are based on:

> *Minimum Description Length Principle in Discriminating Marginal Distributions.* B. Nonchev.
> Pliska Studia Mathematica Bulgarica 22(125), pp. 101-114, 2013.
>
> *Minimum Description1 Length Principle and Distribution Complexity of Spherical Distributions.* B. Nonchev.
> In: Proceedings of the 18th European Young Statisticians Meeting, 2013.
>
> *Minimum Description Length Principle for Fat-Tailed Distributions.* B. Nonchev.
> Proceedings of the 22nd International Conference NDES 2014, Albena, Bulgaria, July 4-6, 2014.
>
> *MDL Principle for Distributions with Shape Parameters.* B. Nonchev.
> Manuscript in preparation.

Contents

Table of Symbols

$L_X(C)$ expected codelength of the code, given that the random variable X is encoded

$A_\epsilon^{(n)}$ typical set

$B_\delta^{(n)}$ high-probability set

F, F_θ a cumulative distribution function, or c.d.f.

H point hypothesis

$H(X)$ The entropy of the random variable X

$L(x^n)$, $L_Bayes(x^n)$, $L_{NML}(x^n)$ universal models, indexes refer to particular universal model

S, S_f, S_X support set of the p.d.f. f or the absolutely continous random variable X

$X \perp Y$ X and Y are independent random variables

$\hat{\theta}$ estimator for θ

$(\Omega, \mathcal{F}, \mathbb{P})$ a probability space with σ-algebra $\mathcal{F}$ and probability function $\mathbb{P}$

$\{X_i\}, \{Y_i\}, \ldots$ a stochastic process in discrete time

$\mathbf{x}^n$ a vector with elements $x_1, x_2, \ldots, x_n$

$\mathcal{C}$, $\mathcal{C}_\aleph$ source code for a random variable having values in $\aleph$

$\mathcal{H}$ statistical model

$\mathcal{R}_{\overline{P}}^{\mathcal{M}}(x^n)$ regret for sample x^n of the distribution $\overline{P}$ with respect of the probabilistic model $\mathcal{M}$

$\mathcal{X}_X$, $\mathcal{X}$ Sample space for the random variable X, defined for any r.v.

$f(x)$, $g(x)$, $f_X(x)$ probability density function (p.d.f.) of a random variable

p, q, p_X probability mass function of a discrete random variable X, i.e. $p_X(x) = \mathbb{P}(X = x)$

a.s. almost surely, i.e. with probability 1

iff if and only if

$\aleph_X$ the values that the random variable X can obtain, also called alphabet in case X is discrete with a finite number of values

$X, Y, \ldots$ Random variables in a suitable probabaility space $(\Omega, \mathcal{F}, \mathbb{P})$. Will be written r.v. for short.

asymptotic equipartition property асимптотична равноделност

Bayesian Information Criteria (BIC) Бейсов информационен критерий

Bayesian universal model / code бейсов универсален модел / код

chain rule верижно правило

codelength дължина на код

distribution complexity сложност на разпределение

entropy, discrete entropy (дискретна) ентропия

Fisher information matrix информационна матрица на Фишер

generalized functions обобщени функции

high-probability set множества с висока вероятност

idealized codelength идеализирана дължина на кода

information criteria информационен критерий (за избор на модел)

Kolmogorov complexity сложност по Колмогоров

Kolmogorov structure function структурна функция по Колмогоров

Kolmogorov sufficient statistic достатъчни статистики по Колмогоров

Minimum Description Length (MDL) principle принцип на минимална дължина на описанието

model complexity сложност на модела

mutual information взаимна информатия

non-singular code несингулярен код

Normalized Maximum Likelihood (NML) model модел на нормализирана максимална правдоподобност

particle swarm optimization оптимизация чрез рояк частици

point hypothesis точкова хипотеза, проста хипотеза

probabilistic model (вероятностен) модел

probabilistic sources вероятностен източник

quantization дискретизация, квантизация

redundancy, maximum redundancy излишество (на кодиране), максимално излишество

regret, minimax regret неефективност (на модел), максимална неефективност

Sequential NML model see *Normalized Maximum Likelihood model*

Shannon code код на Шанън

Shannon-Fano code код на Шанън-Фано

Stochastic Complexity criterion критерий за стохастичната сложност

structure function see *Kolmogorov structure function*

sufficient statistic see *Kolmogorov sufficient statistic*

two-part universal model двустъпков универсален модел

typical set типично множество

uniquely decodable codes еднозначно декодируем код

universal code универсален код

universal model универсален модел

differential entropy диференциална ентропия

joint entropy съвместна ентропия

Kullback-Leibler distance seerelative entropy

Kullback-Leibler divergence seerelative entropy

relative entropy относителна ентропия, дивергенция (дистанция) на Кулбак-Лайблер

source coding кодиране на източник

List of Figures

List of Tables

1. Introduction

A basic task in statistical inference is to find the distribution of a sample, that is, to differentiate between several candidate distributions.

This task is important in a wide variety of applications, and in particular to any method that is sensitive to the assumption that a sample is generated by a random variable with a particular distribution.

Classical non-robust statistical methods often rely on assumptions of normality, which is not warranted in many cases. This is problematic for many reasons, the most general being that samples from a Gaussian distribution exhibit very compact clusters, i.e. the probability of tails events in the sample is severely limited.

Unwarranted assumptions can make parameter estimation inherently unstable.

Moreover, estimation of more complex models, like time-series parameters (e.g. for ARMA-GARCH processes), usually relies on the innovations having a particular distribution. If the residuals calculated are not of the assumed type, goodness of fit and stability of the estimated parameters will suffer, casting doubt on any results obtained with these methods.

A reasonable way to proceed in such cases is to use some kind of robust statistical procedures, that is insensitive (or, up to a certain point, less sensitive) to deviations from distribution assumptions.

This, however, brings its own problems, one of which is that for well-behaved samples robust statistics are generally less efficient than their classical counterparts. In such cases it may be preferable to identify the distribution of a sample and then apply a method specifically targeted for the distribution. Again, we arrive at the task of finding the distribution of a sample.

Another reason for the need to infer the distribution of the sample is for model identification purposes - if the distribution is actually the target of our research. This is a common problem in econometrics, where a cutting-edge research problem for decades has been the identification of stock market returns distribution in the light of the evidence that stock market returns have significantly fatter tails than what we would expect from a Gaussian distribution.[1]

Therefore, in order to talk meaningfully about statistical risk management in econometrics, we have to identify which model is closest to reality.

[1] Also, assymetry and clustering of volatility over time. It is a hard and pressing topic, for which a comprehensive introduction can be foind in [Rachev et al., 2005].

To actually find the distribution of the sample we have to consider a set of possible distributions.

First, we have to posit the alternatives and must make sure that they are reasonable from theoretical standpoint. This is usually done with domain-specific knowledge and is deeply connected with the particular area of application, so how this is done is outside of scope of this dissertation. All we require for our discussion later is that a set of plausible alternatives has been determined.

After we have the set of alternative models, the task becomes to differentiate between them. There are many ways in which that can be achieved, and the use of a particular approach can have deep philosophical implications. Here are some possible methods:

- Kolmogorov-Smirnov type statistics;
- Bayes factors (Example 1.32);
- Kullback-Leibler divergence (i.e. AIC-type information criteria, Example 1.33);
- MDL-based criteria like the Stochastic Complexity criterion (Chapter 2).

All of the approaches have their merits and in certain circumstances can do a decent work.

Kolmogorov-Smirnov (K-S) provides a comparison criteria for a single test. For multiple comparisons, the p-value obtained needs to be adjusted. Moreover, K-S has difficulties with the tails and cannot ensure their closeness, so for tasks where the candidate distributions deviate from each other in the tails, K-S is not a good criterion.

Application of Bayes factors requires prior knowledge, which may not be reasonable, or even clear how to be acquired. It does not take into consideration the flexibility (complexity) of the models, only the goodness of fit, and provides only relative evidence, i.e. which model obtains more support from the evidence.

The AIC-type criteria need to have a reasonably close to the "real" distribution in the considered set, otherwise asymptotics and optimality results do not hold theoretically. Moreover, their simplification can be too great, because many models with different complexity have the same number of parameters. On the other hand they are conceptually simple and have direct connections to the MDL principle.

This dissertation is centered around the Minimum Description Length (MDL) principle and its application to the problem of selecting the distribution of a sample. It is conceptually simple and has deep connection to information theory and in particular Kolmogorov complexity. Moreover it is a practical way to solve applications, and its requirements toward the set of competing models is in some ways much more relaxed.

The main contributions of this dissertation are as follows:

- the introduction of Distribution Complexity (DC)[2], a practical and theoretically sound way to discern between distributions;
- numerical computation mechanisms to calculate the DC for a variety of distributions, including a targeted optimizer implementation.
- calculation of DC for uncorrelated and independent samples for Student-T distribution.

These results have been presented in the following venues, in chronological order:

- Presentation on *"Minimum description length principle in model selection"*, Spring Scientific Session, Section "Probability, Operational Research and Statistics", Faculty of Mathematics and Informatics, Sofia University, 23 March 2012.
- Presentation on *"Minimum Description Length Principle in Model Selection"*, XV International Summer Conference in Probability and Statistics (ISCPS), Pomorie, Bulgaria, 23-30 June 2012.
- Publication titled *Minimum Description Length Principle in Discriminating Marginal Distributions.* B. Nonchev. Pliska Studia Mathematica Bulgarica 22(125), pp. 101-114, 2013. ([Nonchev, 2013b])
- Presentation on *Complexity of spherical distributions using the MDL principle*, Spring Scientific Session, Section "Probability, Operational Research and Statistics", Faculty of Mathematics and Informatics, Sofia University, 16 March 2013.
- Presentation and paper on *Minimum Description Length Principle and Distribution Complexity of Spherical Distributions*, 18th European Young Statisticians Meeting, Osijek, Croatia, 26-30 August 2013. ([Nonchev, 2013a])
- Presentation on *The MDL Principle for Independently Distributed Samples*, Spring Scientific Session, Section "Probability, Operational Research and Statistics", Faculty of Mathematics and Informatics, Sofia University, 29 March 2014.
- Presentation and paper on *Minimum Description Length Principle for Fat-Tailed Distributions.* B. Nonchev. Proceedings of the 22nd International Conference NDES 2014, Albena, Bulgaria, July 4-6, 2014. ([Nonchev, 2014]).
- Presentation on *MDL Principle for Distributions with Shape Parameters,* XVI International Summer Conference in Probability and Statistics (ISCPS), Pomorie, Bulgaria, 21-28 June 2014.

The present dissertation consists of main text split in five chapters and three appendixes with auxiliary material, List of Symbols, List of Figures and List of Tables.

[2]Not related to Complexity distribution, e.g. [Viola, 2010].

The remainder of Chapter 1 introduces the various probability notions that are used throughout this dissertation and elaborates on the challenges of model selection to prepare for the introduction of the MDL principle.

Chapter 2 is a quick tour on the Minimum Description Length principle, including its philosophical implications and connections with information theory and statistics. It is designed to introduce the idea of model complexity, which is the basis for the next chapters.

Chapter 3 describes the contributions of this dissertation to the theory and practice of MDL. It starts with motivation examples of distribution selection, then introduces distribution complexity for spherical distributions, then for independent distributions. The final section is dedicated to the discussion how distribution complexity can incorporate shape parameters (parameters other than scale and location).

Chapter 4 describes the contributions of this dissertation regarding the calculation of the distribution complexity. It presents in details the processes and challenges that arise when calculating the distribution complexity for the various types of distributions. Charts, tables and analysis of the distribution complexity results is also provided.

Chapter 5 concludes the main body of the text with synthesis of the main results and contributions in this dissertation, and outlines future work to extend the distribution complexity to more diverse classes of models.

Additional information on auxiliary concepts and mechanisms is available in the appendices.

Appendix A is a brief guide to the concepts in information theory that justify the connections between information theory and statistics, and in particular the way MDL connects to the classical ideas of Shannon's information theory.

Appendix B describes the main results and properties of generalized functions, which are used in the dissertation as a more succinct way to describe and manipulate integration with respect to conditional distributions.

Appendix C showcases an accompanying custom implementation of a particle swarm optimizer that is used to obtain the numerical results in this dissertation.

1.1. Probability notions

The purpose of this section is to introduce notation and standard definitions that will be used in the rest of the dissertation.

Unless otherwise explicitly specified, the rest of the exposition will assume a suitable probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 1.1. A function $X : \Omega \to \mathbb{R}$ is called *random variable*, or *r.v.*, if it is a measurable function, i.e.

$$\{\omega : X(\omega) \in B\} \in \mathcal{F}$$

for any Borel set B.

The function F_X defined as

$$F_X(x) = \mathbb{P}(X \leq x)$$

is called the *cumulative distribution function*, or c.d.f.

If F_X has a derivative with respect to x for all $x \in \mathbb{R}$ then X is called *absolutely continuous* and

$$f_X(x) = \frac{\partial F_X}{\partial x}(x)$$

is called the *probability density function*, or p.d.f.

If X is discrete, i.e. has a countable number of values, then its c.d.f. is a stepwise function, so it does not have a derivative. In this case

$$p_X(x_i) = \mathbb{P}(X = x_i)$$

is called the *probability mass function*, or p.m.f.

In all cases above it is possible to omit X, if it is evident from the context.

Definition 1.2 (Convergence of random variables)**.** Given a sequence of random variables $X_1, X_2, \ldots$, there are several ways in which we can say it converges to a random variable X

1. in probability, if for every $\epsilon > 0$, $\mathbb{P}\{|X_n - X| > \epsilon\} \to 0$;
2. in mean square if $\mathbb{E}(X_n - X)^2 \to 0$;
3. with probability 1 (*almost surely,* or a.s.), if $\mathbb{P}\{\lim_{n\to\infty} X_n = X\} = 1$.

Definition 1.3 (Stochastic process)**.** Let $\{X_t\}_{t\in T}$ be a set of random variables indexed by a parameter $t \in T$. They form a random process if their index can be interpreted as a logical ordering of the variables, e.g. through time.

If $T \subseteq \mathbb{Z}$ then $\{X_i\}_{i\in T}$ is called a process in discrete time.

The defining property of stochastic processes are their joint c.d.f.. If X_t are absolutely continuous variables, their joint p.d.f. can be used. For discrete-value X_t, the joint p.m.f. is used.

Markov processes are a special class of processes whose future depends only on the current state.

Definition 1.4 (Markov process)**.** The stochastic process $\{X_t\}$ is called *Markov process*, if its future is conditionally independent of its past given the present, i.e.

$$\mathbb{P}(X_{t_n} \leq x_{t_n} | X_{t_{n-1}} \leq x_{t_{n-1}}, \dots X_{t_1} \leq x_{t_1}) = \mathbb{P}(X_{t_n} \leq x_{t_n} | X_{t_{n-1}} \leq x_{t_{n-1}})$$

for all $t_1 \leq t_2 \leq \ldots \leq t_n$.

Markov processes in discrete time are called *Markov chains*.

Definition 1.5 (Markov chain)**.** The random variables X, Y and Z are said to form a Markov chain in that order (denoted by $X \to Y \to Z$) if the conditional distribution of Z is conditionally independent of X, i.e.

$$p_{X,Y,Z} = p_X p_{Y|X} p_{Z|Y}.$$

1.1.1. Estimation

The most common task in statistical inference is to pick the exact distribution from a set of alternatives. When the alternatives are parameterized via a parameter $\theta \in \Theta$, this task is more commonly called *estimation*.

Definition 1.6 (Statistic)**.** Let x^n be a sample. Any function of the sample $T : S_X^n \to \mathbb{R}^k$ is called a *statistic*.

Definition 1.7 (Estimator)**.** Let $\mathcal{M} = \{F_\theta\}$ be a family of distributions, parameterized by a (vector) parameter $\theta \in \Theta$. A statistic$T(x^n)$ with values from Θ, which is used to infer the value of the parameter that produced the sample, is called an *estimator*.

It is usually denoted by $\hat{\theta}$.

Good estimators are expected to have the following properties:

Definition 1.8 (Bias)**.** Let $T(x^n)$ be an estimator of $\theta \in \Theta$. The *bias* of an estimator $T(X_1, \dots, X_n)$ for the parameter θ is the difference between the expected value of the statistic and the real value of the parameter:

$$bias(T) = \mathbb{E}_\theta T(X_1, X_2, \dots, X_n) - \theta.$$

The estimator T is called *unbiased*, if for all $\theta \in \Theta$ we have

$$\mathbb{E}_\theta T(X^n) = \theta,$$

where the subscript θ in $\mathbb{E}_\theta$ means that the r.v. X is distributed according to F_θ.

The subscript signifies that the expectation is with respect to the density $f_\theta(x)$.

Bias tells only part of the story, however. We also need to look at the magnitude of the error, which is often evaluated in terms of a mean square error.

Definition 1.9 (Variance)**.** Let $T(x^n)$ be an estimator of $\theta \in \Theta$. Its variance is defined as

$$var(\hat{\theta}) = \mathbb{E}_\theta \left(T(X^n) - \theta\right)^2 .$$

Naturally, the smaller the variance, the better the estimator.

Definition 1.10. An estimator $T_1(X_1, \ldots, X_n)$ is said to *dominate* another estimator $T_2(X_1, \ldots, X_n)$ if, for all θ,

$$\mathbb{E}(T_1(X_1, \ldots, X_n) - \theta)^2 \leq \mathbb{E}(T_2(X_1, \ldots, X_n) - \theta)^2.$$

The variance of unbiased estimators is limited from below by the following theorem.

Theorem 1.11 (Cramér-Rao bound, Theorem A.72)**.** *Let $\hat{\theta}(x^n)$ be an unbiased estimator for $\theta \in \Theta$. Then*

$$var(\hat{\theta}) \geq [I(\theta)]^{-1} , \tag{1.1}$$

where $I(\theta)$ is the Fisher information.

Fisher information is introduced in Definition A.71 in Section A.5.4.

For some parameters, the variance of the unbiased estimators can actually achieve the limit from Theorem A.72.

Definition 1.12 (Efficiency)**.** If an unbiased estimator $\hat{\theta}$ satisfies (1.1) with equality, it is called *efficient*.

Definition 1.13 (Consistency)**.** Let $T(x^n)$ be an estimator of $\theta \in \Theta$. $T(X_1, \ldots, X_n)$ is called *consistent* estimator for θ if

$$T(X_1, \ldots, X_n) \to \theta \text{ in probability as } n \to \infty.$$

Consistency is desired from theoretical standpoint, because even though in practice we often have very limited data, we want to be certain that if more data is obtained, the estimated parameter value will converge to the true parameter value.

Note that consistency, bias and variance all refer circularly to the distribution in question, e.g. the expected value in the bias calculation refers to the true value of the parameter. If the process generating the data is not described well by any member of the parametric family $\mathcal{M}_\theta$, then unbiasedness does not guarantee that the expected value of the statistic will equal any $\theta \in \Theta$.

A related question is raised in Section 1.1.1.3 regarding the meaning of bias in terms of the estimated value of the parameter.

Finally, we introduce the sample estimators. They are central in the analysis of the complexity of spherical distributions in Chapter 3.

Definition 1.14 (Sample mean and variance)**.** Let $x^n = (x_1, x_2, \ldots, x_n)$ be a sample. The *sample mean*, denoted by $\overline{x}$, is defined by

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The *sample standard deviation*, denoted by s_x, is defined as

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 .$$

1.1.1.1. Likelihood and the Likelihood Principle

Let $\theta \in \Theta$ be a parameterization of a family of distributions $\mathcal{M}_\theta$.

Definition 1.15 (Likelihood)**.** Let $p(x^n|\theta)$ be a family of p.m.f., indexed by a parameter $\theta \in \Theta$. Considering p as a function of θ, $p(x^n|\theta)$ is called the *likelihood* of θ given a sample x^n.

Let $f(x^n|\theta)$ be a family of distribution p.d.f.s indexed by a parameter θ. Considering f as a function of θ, $f(x^n|\theta)$ is called the *likelihood* of θ given a sample x^n.

The likelihood represents how well the given value of θ fits a sample x^n.

For mathematical simplicity it is often the log-likelihood $\log p(x^n|\theta)$ or $\log f(x^n|\theta)$ that is used instead. Moreover, the log-likelihood is connected to the description length of the data (Section 2.3.4).

A controversial principle in statistics is the *likelihood principle* that states that all the information contained in a sample x^n is contained within the values of the likelihood function. Many analyses, like calculation of classical confidence intervals and p-values, do *not* observe the likelihood principle and depend significantly on the design of the experiment, i.e. not only what happened, but what could have happened.

One way to apply the likelihood principle is by *maximum likelihood estimation* (Section 1.1.1.2) and *Bayesian inference* (Section 1.1.1.3 and Example 1.32). The concept of *sufficient statistics* (Section 1.1.2) also relies on the likelihood principle.

1.1.1.2. Maximum Likelihood Estimation

A common method to determine the value of a parameter θ is to use the likelihood as the measure of how well a given parameter value fits the sample.

Definition 1.16 (Maximum Likelihood Estimator (MLE))**.** The estimator $\hat{\theta}$ of $\theta \in \Theta$ defined as

$$\hat{\theta}(x^n) = \arg\max_{\theta} f(x^n|\theta)$$

is called the *maximum likelihood estimator*.

MLEs are widely used, because of its conceptual simplicity and attractive properties.

1. Under very lax regularity conditions, a MLE is guaranteed to be consistent.
2. Likelihood is a proxy of goodness of fit measure, so MLE is in a sense the best-fit estimate.

Because MLE takes into consideration only how well the distribution $f(x|\theta)$ fits the data, having more parameters to fit will always lead to lower likelihood. This may be caused by genuinely better fit, or just random fluctuations in the data. This line of reasoning is further explored in Section 1.3.

1.1.1.3. Bayesian inference

Another popular way to find estimators is to consider $\theta \in \Theta$ as a random variable, and incorporate the information from x^n using Bayes formula for conditional probability.

Definition 1.17 (Prior distribution)**.** If $\theta \in \Theta$ is a parameter of a family of distributions $\mathcal{M}$, a prior distribution is called any distribution $\pi(\theta)$ over Θ (p.d.f. or p.m.f.).

Prior distributions incorporate the apriori knowledge of the parameter θ.

Definition 1.18 (Posterior distribution)**.** Let $\theta \in \Theta$ be a parameter of a family of distributions $\mathcal{M}$, with assigned prior distribution $\pi(\theta)$. The *posterior distribution* is derived using Bayes theorem as

$$p(\theta|x^n) = \frac{p(x^n|\theta)\pi(\theta)}{\int_{T\in\Theta} p(x^n|T)\pi(T)dT} = \frac{p(x^n|\theta)\pi(\theta)}{p(x^n)}. \tag{1.2}$$

Since the apriori probability $p(x^n)$ is not dependent on the parameter, we can safely ignore it and rewrite (1.2) as

$$p(\theta|x^n) \propto p(x^n|\theta)\pi(\theta).$$

Using the posterior distribution we can derive different point estimates for θ - using expectation, median, mode of $p(\theta|x^n)$, or use the more general interval estimates.

Bayesian analysis, contrary to classical (also called *frequentist*) statistics does not need to create circuitous arguments in order to answer questions about the parameter θ because they are defined as random variables.

Another point is that the bias, introduced in Section 1.1.1, guarantees us that the expected value of the statistic, given the parameter, is equal to it, i.e. (with a slight abuse of notation)

$$\mathbb{E}(X|\theta) = \theta.$$

What we really want to know is that the expected value of the parameter is the value of the statistic:

$$\mathbb{E}(\theta|X) = X.$$

Lack of bias does not guarantee that, for example see [Blume, 1975]. In the Bayesian paradigm these kinds of questions are tractable.

1.1.2. Sufficient statistics

Sufficient statistics were introduced by Fisher to isolate the propensity of some distributions, namely exponential families, to concentrate all the available information for a sample into a simple statistic.

The notion of sufficient statistic is used in Section 3.1.2 to provide another way of looking at the model complexity, that in case of scale-location families leads to the novel concept of distribution complexity (Chapter 3).

Theorem 1.19 (Fisher-Neyman factorization)**.** *Let $f_\theta(x^n)$ be a family of distributions, parameterized by θ. $T(x^n)$ is a sufficient statistic of θ if and only if non-negative functions $h(x^n)$ and $g_\theta(T)$ can be found, such that*

$$f_\theta(x^n) = h(x^n)g_\theta(T(x^n)) \tag{1.3}$$

Another way to look at sufficient statistic is via the alternative expression

$$f(\theta|T(x^n) = t, x^n) = f(\theta|T(x^n) = t),$$

which means that the conditional likelihood of the parameter θ, given the sufficient statistic $T(x^n)$, does not depend on the specific x^n that gave rise to it.

Proof. See [Kay, 1993], Appendix 5A. □

Note that Theorem 1.19 works for both continuous and discrete random variables, with f being the p.d.f. for the continuous case and f being the p.m.f. for the discrete case.

Theorem 1.20. *If $f_\theta(x^n)$ is a family of distributions and $T(x^n)$ is a sufficient statistic of θ, then in (1.3) the function $g_\theta(T(x^n))$ can be chosen to be the probability distribution of $T(X^n)$.*

In this case

$$h(x^n) = f^\theta_{X^n|T(X^n)=t}(x^n)$$

is the conditional probability distribution of X^n on the set $x^n \in B(t_0)$, where

$$B(t) = \{x^n \in \mathcal{X}^n : T(x^n) = t\}.$$

Proof. Fix a $t_0 \in \{t : B(t) \text{ is not empty}\}$. On one hand

$$f_\theta(x^n) = f^\theta_{T(X^n)}(T(x^n)) f^\theta_{X^n|T(X^n)=t}(x^n),$$

while due to Theorem 1.19, we can also decompose it as

$$f_\theta(x^n) = h(x^n) g_\theta(T(x^n)).$$

From the two decompositions it is clear that

$$f^\theta_{X^n|T(X^n)=t}(x^n) = \frac{g_\theta(T(x^n))}{f^\theta_{T(X^n)}(T(x^n))} h(x^n) = \frac{g_\theta(t)}{f^\theta_{T(X^n)}(t)} h(x^n). \tag{1.4}$$

Define a normalization term

$$s(t) = \frac{g_\theta(t)}{f^\theta_{T(X^n)}(t)}$$

and define new decomposition functions $h^*(x^n)$ and $g^*_\theta(t)$ as

$$g^*_\theta(t) = g_\theta(t) s(t)$$

$$h^*(x^n) = \frac{h(x^n)}{s(t)}.$$

Using the new decomposition functions in (1.4) leads to the required decomposition. □

The following corollary uses δ-calculus (see Appendix B) for easier technical manipulation of conditional distributions further down the text.

Corollary 1.21. *Using the decomposition in Theorem 1.20 we have for all t in the domain of $T(X^n)$ that*

$$\int h(x^n) \delta(T(x^n) - t) dx^n = 1.$$

Proof. Consider the following equation chain

$$
\begin{aligned}
\int h(x^n)\delta(T(x^n)-t)dx^n &= \int f^\theta_{X^n|T(X^n)=t}(x^n)\delta(T(x^n)-t)dx^n \\
&= \int \frac{f^\theta_{X^n,T(X^n)}(x^n,t)}{f_T(t)}\delta(T(x^n)-t)dx^n \qquad (1.5)\\
&= \int \frac{f^\theta_{X^n}(x^n)\mathbf{1}_{T(x^n)=t}(x^n,t)}{f_T(T(x^n))}\delta(T(x^n)-t)dx^n \\
&= \int \frac{\mathbf{1}_{T(x^n)=t}(x^n,t)}{f_T(T(x^n))}\delta(T(x^n)-t)dF^\theta_{X^n}(x^n) \\
&= \int \frac{\delta(T(x^n)-t)}{f_T(T(x^n))}dF^\theta_{X^n}(x^n) \\
&= \int \frac{\delta(s-t)}{f_T(s)}dF^\theta_{T(X^n)}(s) \qquad (1.6)\\
&= \int \delta(s-t)\frac{f_T(s)}{f_T(s)}ds \\
&= 1,
\end{aligned}
$$

where (1.5) follows from the definition of conditional distribution, and (1.6) is trivial, because the right side no longer depends on the actual value of x^n, but only on the statistic $T(x^n)$, so we changed the integration to the transformed variable $S = T(X^n)$. □

The following example provides explicit formulas for the sufficient statistics for a Gaussian family and is used later in Section 3.1.2 as the motivating example.

Fact 1.22 (Sufficiency in Gaussian family)**.** *Let x^n be an i.i.d. sample of Gaussian distributed random variables from $N(\mu,\sigma^2)$. The sample mean $\overline{x} = \frac{1}{n}\sum_i x_i$ is sufficient statistic for μ.*

Sufficient statistics for μ and σ are $\overline{x}$ and the sample standard deviation $s_x = \sqrt{\frac{1}{n}\sum(x_i-\overline{x})^2}$ together.

Proof. The joint density of the sample is expressible as

$$
\begin{aligned}
f(x^n|\mu,\sigma) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{ns_x^2}{2\sigma^2}\right)\exp\left(-\frac{(\overline{x}-\mu)^2}{2\left(\frac{\sigma}{\sqrt{n}}\right)^2}\right). \qquad (1.7)
\end{aligned}
$$

Since all terms in (1.7) that are dependent on x_i are s_x and $\overline{x}$, the factorization (1.3) is obvious. □

Note that the Pitman–Koopman–Darmois theorem requires that for i.i.d. samples the only distributions that have bounded sufficient statistics as sample size increases form exponential families, suggesting that the concept of sufficiency is of limited value.

This however is not necessarily true when the samples are *not* i.i.d., in particular when the independence assumption is not strictly satisfied - as is the case with multivariate Student-T distribution, where only linear independence is provided (i.e. zero correlation). This is further used in Section 3.3 to calculate the complexity of spherical distributions.

1.1.3. Quantization of random variables

A lot of ink has been spent on the difference in notation between discrete and continuous random variables. When it comes to information theory this distinction is most artificial, because eventually all continuous random variables are discretized in order for a measurement to be written down. Consequently, it is beneficial to use a common basis for both discrete and continuous variables.

One straightforward idea to expand methods and algorithms designed over discrete random variables to continuous random variables is to define a *quantization*. It is a scheme of splitting the values of the continuous variable so that the resulting ranges have intuitive interpretation.

Quantization is also useful when the algorithms and methods have intuitive explanation for discrete random variables. There are many such possible schemes, in this section we introduce those that are useful for the purposes of information theory.

Definition 1.23 (Partition)**.** Let $\mathcal{X}$ be the range of values of the random variable X. A partition $\mathcal{P}$ of $\mathcal{X}$ is a finite collection of disjoint sets P_i such that $\cup_i P_i = \mathcal{X}$.

Definition 1.24 (Quantization)**.** The *quantization* of X by $\mathcal{P}$ is denoted by $[X]_{\mathcal{P}}$ is a discrete random variable defined as

$$\mathbb{P}\left([X]_{\mathcal{P}} = i\right) = \mathbb{P}\left(X \in P_i\right) = \int_{P_i} dF(x).$$

Lemma 1.25. *Let X be a random variable and $\mathcal{P}_n$ be a sequence of partitions with decreasing size and let $g(x)$ be Lebesgue integrable under the density of X (regularity condition). Then*

$$\int g(x) dF_{[X]_{\mathcal{P}_n}}(x) \to \int g(x) dF_X(x).$$

From the above lemma we have that expectation, standard deviation, correlation and other relations between the quantized versions converge to the values for the original variables with decreasing partition size.

Next we introduce the particular quantization scheme that we will use.

Definition 1.26 (Quantization scheme)**.** Let X be a random variable and let $x_i = i\Delta$ be the middle points of the *partition* intervals.

By the mean value theorem, it is always possible to pick $x_i^\alpha \in \left(\left(i - \frac{1}{2}\right)\Delta; \left(i + \frac{1}{2}\right)\Delta\right)$ such that

$$\int_{\left(i-\frac{1}{2}\right)\Delta}^{\left(i+\frac{1}{2}\right)\Delta} f(x)dx = f(x_i^\alpha)\Delta.$$

The quantized version of X is the discrete random variable X^Δ is defined as

$$X^\Delta = x_i^\alpha \text{ whenever } \left(i - \frac{1}{2}\right)\Delta \leq X < \left(i + \frac{1}{2}\right).$$

The p.m.f. for X^Δ is then defined as

$$p_{X^\Delta}(x_i) = \mathbb{P}\left(X \in \left[x - \frac{\Delta}{2}; x + \frac{\Delta}{2}\right)\right) = \int_{\left(i-\frac{1}{2}\right)\Delta}^{\left(i+\frac{1}{2}\right)\Delta} f(x)dx = f(x_i^\alpha)\Delta,$$

i.e. the p.m.f. of X^Δ is proportional to the p.d.f. of X.

The type of quantization scheme presented in Definition 1.26 naturally preserves (asymptotically by partition size) the properties of the random variable that are important in information theory.

Additionally, the quantization intervals are of equal size, which is reasonable - when we talk of quantization we usually imply partitioning the domain into equally-sized chunks. The alternative, an equal-probability split, has several drawbacks:

1. It is dependent on the measure of probability *and* the distribution of X, which we do not always want to fix, particularly in the setting of model selection.
2. For information-theoretic purposes, the "information" that we want to measure is often contained in the probability distributions. Thus fixing a probability distribution changes the information content of the quantized random variable, which is clearly not desirable.

For those reasons the quantization scheme in Definition 1.26 is intuitive and preferable.

1.2. Information theory

This section introduces some basic definitions from Shannon's information theory. Further details are provided in Appendix A.

First, the most basic quantity of a random variable in information theory is its entropy. It is defined as follows:

Definition 1.27 (Definition A.2)**.** The *entropy* of a discrete random variable is defined as

$$H(X) = -\sum_{x \in \aleph_X} p(x) \log p(x) = -\mathbb{E}_X \log p(X), \tag{1.8}$$

For continuous variables, the definition is extended to the following to take into consideration the scale of the variable as well:

Definition 1.28 (Definition A.5)**.** Let X be an absolutely continuous random variable with p.d.f. $f(x)$. Its *differential entropy* is defined as

$$h(x) = -\int_{S_X} f(x) \log f(x) dx.$$

Entropy represents the uncertainty of the variable's value. If the uncertainty is low, the entropy is also low. Random variables that are constant have discrete entropy of 0, and differential entropy of $-\infty$. The discrete and continuous entropy are related via the quantization from Section 1.1.3, which is shown in detail in Section A.2.3.

A related quantity that is used in statistics is the *relative entropy*, or *Kullback-Leibler divergence*.

Definition 1.29 (Definition A.18)**.** If $p(x)$ and $q(x)$ are two probability mass functions, the Kullback-Leibler distance between them is defined as

$$D(p \parallel q) = \sum_{x \in \aleph} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \log \frac{p(X)}{q(X)}.$$

It is finite only when $q(X) > 0$ for all x for which $p(x) > 0$.

The Kullback-Leibler divergence $D(f \parallel g)$ between two densities f and g is defined by

$$D(f \parallel g) = \int f \log \frac{f}{g} dx.$$

Note that $D(f \parallel g)$ is finite only if $S_f \subseteq S_g$.

Kullback-Leibler divergence can be described as a "distance" between the real distribution p (or f) and another distribution q (or g). It is not symmetric, thus is not a measure, but is the basis of Akaike's AIC and the other so-called information criteria (Example 1.33).

In order to do inference based on encoding a sample, the sample has to be encoded in a D-ary alphabet. An optimal code to do that is the Shannon code, for which we have the following theorem considering the expected codelength:

Theorem 1.30 (Shannon code, Theorem A.56)**.** *Let X be a random variable. There is code with lengths assigned as*

$$l_i = \lceil -\log_D p_i \rceil$$

that satisfies the bounds

$$H_D(X) \leq L < H_D(X) + 1,$$

called Shannon code*.*

1.3. Model selection

This section shows the basic notions regarding model selection in statistics. It is centered over the goodness of fit/generalizability duality and ends with a brief introduction of the so-called information criteria AIC, BIC, DIC, etc. and the introduction of what is often refered to as the MDL criterion.

First we start with the basic definition of models.

Definition 1.31 (Model)**.** A family of related distributions $\mathcal{M}$ is called a *model*.

Each model is usually parameterized by some (multivariate) parameter $\theta \in \Theta$. A common task when presented with a (possibly infinite) list of models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ is to find the one that "best" represents the observed process. Inference is done on a realized sample of the process.

There are two basic requirements for a model in order to be considered good:

- goodness of fit
- generalizability

The task of model selection is to find the "best" model, considering these two (often contradicting) qualities. Their definitions and specifics are further explored in the following subsections.

1.3.1. Goodness of fit

A good model should fit (explain) past data well. This is an obvious requirement. We cannot expect to explain future data well if we don't make sense of the past.

The criteria of the good fit, or more precisely, which measure of fit to use, is not unambiguous. The goodness of fit measure is often intertwined with the method of fit, or more frequently the method of fit is defined to be the one maximizing the chosen goodness of fit measure.

Classical criteria often use the quadratic norm of the unexplained data (residuals), and those procedures are optimal for Gaussian models, but when non-normality is present they can show very bad fit or even misleading results. Their main advantage is computational simplicity.

Robust measures of fit are often harder to construct and analyze, but handle well errors in data. Their drawback is their lower efficiency at homing at the correct model, requiring more data to estimate the parameters with reliability when the sample comes from a Gaussian.

1.3.2. Model Generalizability

Generalizability is the ability of the model to explain future data. It is a more elusive quality, but very important for several reasons.

First, data is subject to experimental error and noise. This necessitates adequate sensitivity to input data, i.e. the estimated models should have stability when it comes to the small changes in the data.

Second, sample data is often limited. Use of statistical models on a limited dataset requires confidence that future datasets will have similar properties, and the largest common denominator of those properties is not likely to contain all properties seen in a small sample.

In general, having a too finely tuned model to explain a dataset can drastically reduce the performance of the model for future data. The problem is exacerbated by the abstractness of future data, and the key is the *model complexity*, which will be discussed in Section 2.6.

1.3.3. Model identification

A related task to model selection is model *identification*. This is the process of using the best model found to draw conclusions on the nature of the process that has generated it. Informally, if it quacks like a duck and it walks like a duck, it is a duck. Or is it?

Model identification is rather more involved with the philosophy of statistical analysis and statistical model selection. Paradigm, like Bayesian inference (Section 1.1.1.3) posit the existence of a "true" model, which although random in nature is still "real". MDL as a philosophy attaches rather less importance on the existence of real mechanism (more on that in Section 2.3) and focuses instead on the data.

In general model identification is not a purely statistical task, and is more connected to the object of statistical analysis rather than the statistical analysis itself. Consequently, most statistical constructs do a poor job of handling model identification and instead relegate this aspect of statistical modeling to the field of the studied process.

1.3.4. Examples

In summary, these goodness of fit and generalizability often give contradictory indications when several competing models are compared. A more complex model with many variables may easily fit any data better than a more parsimonious one, but if the fitted parameters try to explain the noise, instead of the underlying relationship (i.e. suffer overfit), they can easily provide poorer explanation of future data, so in turn generalizability would suffer.

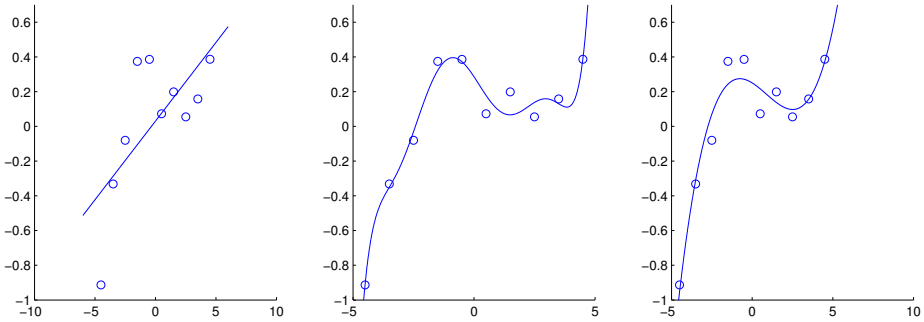

Figure 1.1.: Fit vs Generalizability simulation. The same data fitted with linear model (left), 7th-degree polynomial (center) and 3rd degree polynomial (right). . Real data is generated with 3rd degree polynomial.

These concepts are illustrated in Figure 1.1. The goal of model selection in this case is to find the proper degree of the polynomial.

Clearly, the model on the left is not adequate, as what is deemed noise is quite large, and moreover, that noise has clearly visible structure. Its simplicity, however, provides the advantage that interpolation and extrapolations are very robust and do not change easily - so it has high generalizability.

The center model uses 7-degree, and while it follows the data very closely, there are some aspects that clearly do not make sense. For example some kinks in the fitted line that are way too specific to be derived by this small amount of data. Moreover, if we want to extrapolate, the forecast line curves quickly upwards, which is also not warranted by the data. This means poor generalizability.

We can say that given the data, the model on the right is the "proper" one, because it fits the data reasonably well, and generalizes well enough - strikes a balance. To quantify that balance, various criteria have been developed over the years.

Note that the same considerations apply as well even when no "correct" model is provided and we just want to make sense of the data.

Some examples how different model selection criteria find the balance between generalizability and goodness of fit follow below.

Example 1.32 (Bayes factors)**.** A Bayesian approach to model selection, alternative and in a way similar to classical hypothesis testing, is the Bayes factors approach of [Kass and Raftery, 1995].

Let M_1 and M_2 are two candidate models, parameterized over Θ_1 and Θ_2. The Bayes factors approach consists of calculating the ratio K of the posterior probabilities of the two models:

$$K = \frac{p(x^n|M_1)}{p(x^n|M_2)} = \frac{\int_{T_1 \in \Theta_1} p(x^n|T_1, M_1)p(T_1|M_1)dT_1}{\int_{T_2 \in \Theta_2} p(x^n|T_2, M_2)p(T_2|M_2)dT_2}. \tag{1.9}$$

The interpretation is that if $K > 1$ then the data provides evidence towards the hypothesis M_1 over M_2.

Note that equation does not take into consideration (or assign) a prior probability of which model is correct, M_1 or M_2. This means that the evidence obtained is on the relative merit of the two models - if M_1 is very unlikely on other grounds, then even high K may not make it the preferred explanation. This aspect is not incorporated in (1.9).

There is, however, a kind of built-in compensatory mechanism for overfitting in (1.9) - if one of the models is very complex, that is, if it explains a lot of data with high goodness of fit, then the probability of the data given the model would be smaller, hence would provide less support for this model.

This is behavior similar to the model complexity in MDL, see Section 2.6.2.

Now we will present the criteria for model selection that derive from some form of bias correction, and as a consequence reduce to the minimization of the sum of a goodness of fit term and a penalization term.

Example 1.33 (AIC, DIC, GIC)**.** The so-called information criteria that started exploding in popularity since their inception in [Akaike, 1974] are based on the idea of finding the model that is closest to the data. That is achieved by minimizing the Kullback-Leibler divergence, which is introduced in Definition A.18.

In the end, the analysis is reduced to finding the model that maximizes the "true" likelihood of the given data.

Let $\mathcal{M}_\theta$ be a family of distributions and $\log f(x^n|\theta)$ be the log-likelihood of the parameters. Let also $\hat{\theta} \in \Theta$ be the MLE, defined by

$$\hat{\theta} = \arg\max_\theta \log f(x^n|\theta).$$

What we want to do, in addition to getting the value $\hat{\theta}$, is to obtain also the goodness of fit, and the natural choice is to estimate $\log f(x^n|\theta)$ with $\log f(x^n|\hat{\theta})$, the "observed" value of the goodness of fit.

However, this value is biased upwards, because in order to estimate it, we first use the data x^n to estimate $\hat{\theta}(x^n)$ and then use that estimate again with x^n to estimate the likelihood.

Earlier solutions to remove the bias were the various bootstrap methods, where part of the sample was used for fitting the data and the rest to estimate the goodness of fit.

Akaike's idea in [Akaike, 1974] was to actually estimate the bias, and correct for it. It turns out that for models satisfying some mild regularity conditions,

$$bias(\log f(x^n|\hat{\theta})) \longrightarrow k,$$

where k is the number of free parameters (the dimensionality of θ) and the asymptotics is by sample size.

Thus AIC discriminates between models by finding the highest minimum value of

$$AIC = -2\log f(x^n|\hat{\theta}) + 2k.$$

GIC and DIC are derived by the same basic idea, but cover other methods of fit than MLE. A comprehensive introduction on the topic of information criteria, their motivation, derivation and properties can be found in [Konishi and Kitagawa, 2008].

Example 1.34 (BIC)**.** The so-called Bayesian information criterion (BIC), introduced by Shwarz in [Schwarz, 1978] is not related to Kullback-Leibler divergence, so its designation as information criterion (IC) is a misnomer.

BIC is based on finding the maximum posterior probability of the model. As in Example 1.33, this probability can be estimated by function of x^n and $\hat{\theta}(x^n)$, so it must be corrected for bias. This bias is also derived asymptotically as in AIC.

The BIC is applied by minimization of

$$BIC = -2\log f(x^n|\hat{\theta}) + k\ln N,$$

i.e. with a penalty term $k\ln N$ instead of $2k$.

Example 1.35 (MDL)**.** Another criterion to find the best model is the MDL criterion, which follows the MDL principle that will be introduced in details in Chapter 2.

Broadly speaking, the goodness of fit is estimated using the length of the description of the data (which for many models coincides with the negative log-likelihood, up to an additive constant).

Generalizability is then quantified by the description length of the model. The criterion then takes the form of minimizing

$$MDL = L(x^n|\theta) + L(\theta), \tag{1.10}$$

which, using idealized codelength from Definition 2.11 as

$$L(x^n|\theta) = -\log f(x^n|\hat{\theta})$$

reduces (1.10) to

$$MDL = -\log f(x^n|\hat{\theta}) + L(\theta).$$

Particular implementations of this criterion use different description lengths of the hypothesis. For example, the Stochastic Complexity criterion (see Section 2.6.3) use the model complexity as description length.

This criterion and the related MDL principle is further discussed in Chapter 2.

2. The Minimum Description Length Principle and Applications

This section serves as an introduction to the ideas behind the Minimum Description Length principle and its philosophical implications. It contains the classical results in the area and their connection with principles from Shannon's information theory presented in Appendix A and Section A.5 and serves to put into perspective the concept of distribution complexity from Chapter 3.

For a thorough treatment of this vast topic the reader is referred to [Grunwald, 2007]. A more statistics-oriented approach is available in [Rissanen, 2007], focusing on the complexity of the models and its application in statistical inference.

The contents of this chapter are organized in the following way.

First we motivate the MDL principle with classical examples in Section 2.1 and the connection to information theory, then we introduce the basic notions in Section 2.2. The philosophy behind the MDL principle is covered in Section 2.3.

Section 2.4 introduces the Crude MDL principle, and Section 2.5 builds upon it to define the Refined MDL principle, and consequently introduce the notion of universal model.

In Section 2.6 the Normalized Maximum Likelihood model, model complexity and Stochastic Complexity criterion are presented. They form the basis for the distribution complexity in Chapter 3, the main contribtion of this dissertation.

We conclude with Section 2.7, containing a brief review of recent advances in the theory and application of the MDL principle.

2.1. Introduction

The ultimate goal of model selection is to find models that are useful in understanding the real world. The question whether they are "true" is not always relevant or tractable.

To motivate the use of the MDL principle we begin by first casting the motivation behind Kolmogorov complexity from Section A.6 into a more statistical inference-oriented viewpoint.

The task is to transmit a string with million symbols 0 or 1 each. Let's consider he following three example sequences:

- 0101010101010101010101010101010101010...
- 1101100111111101111110110011111111111...
- 1010101000111010001110100011101011110...

On one hand, there are better ways of transmitting the whole sequence, other than bit by bit, if we exploit regularities in the data. On the other, regularities in the data suggest what should be the algorithms used to send the sequences.

The first sequence is just 01 repeated, so sending this instruction instead of the sequence is quite a lot faster. The second sequence has about 9 ones for each zero, so we can encode long strings of ones with shorter codes, thus transmit shorter instructions (on average). For the third sequence there is not much we can do, as it is (ostensibly) random and uniform, with equal probabilities of 0 and 1, and no discernible pattern of repetitions are available to exploit.

In each of the cases, knowledge of the patterns in the data allows us to compress it.

However, using the most general case of descriptions, a Turing-complete machine (universal computer) and arbitrary program, as explored in Section A.6, results in codes that are not computable, and are, to some extent, arbitrary in length - dependent on the exact computer.

The main insight of Rissanen for the MDL principle is to restrict the set of codes to those corresponding to probability distributions. When doing it, we assume only that the probability distribution is a description method - we do not posit that the modeled sample is really generated by a random process.

If we have a discrete random process X with p.m.f. $p_X(x)$, and observe a sequence of n realizations, by applying a Shannon code (Theorem A.56) we can obtain code C for X with codelength

$$l(x^n) = \lceil -\log p_X(x^n) \rceil .$$

For an absolutely continuous random process Y with p.d.f. $f_Y(x)$, we can do a quantization as in Section 1.1.3 with pre-determined partition size Δ. The ideas from Section 1.1.3 allows us to define a code C that will encode n realizations of Y with codelength

$$l(x^n) = -\log f_y(x^n) + \Delta.$$

In both cases we will have a correspondence between codes and distributions, and we will define the so-called *idealized codelength*:

$$l(x^n) = -\log P(x^n), \tag{2.1}$$

up to a constant and for a suitable $P(x^n)$.

With (2.1) in mind, when evaluating a set of models $\mathcal{M}_\theta$, the codelength that is achieved when a particular θ is equal to the negative of the log-likelihood. This is useful because codelength is a more natural measure for describing θ, and we can combine the goodness of fit and generalizability into a single measure. This is explored further in Section 2.3.

In the literature a set of distributions with some defining characteristic (e.g. the set of all Gaussian distributions) is called a *model*. How inference is done to choose the appropriate model via the MDL principle is further explored in Section 2.2.

For each distribution there is an optimal code, namely the Shannon-Fano code. A single distribution that approximates the codelength for all distributions in a model "well" is called an *universal model* and the main line of research on the MDL principle is the discovery of those models. They are introduced in Section 2.5.

2.2. Notions

In order for the MDL principle to offer a unifying framework for the treatment of discrete and continuous variables alike, we will consider the following definitions for *probabilistic sources* and *probabilistic models.*

Definition 2.1. The values that the random variable X can take is called the *sample space* of the r.v. and is denoted with $\mathcal{X}_X$.

When the random variable is implied, the index can be skipped as $\mathcal{X}$.

In most practical applications the random variables are either discrete or absolutely continuous, so the notion of *sample space* is used as a substitute for either *alphabet* from Section A.2.1 or *support set* from Section A.2.2.[1]

From Definition 2.1 we can extend the concept of sample space to the trivial case:

$$\mathcal{X}^0 = \left\{x^0\right\},$$

the multivariate case:

$$\mathcal{X}^n = \underbrace{\mathcal{X} \times \mathcal{X} \times \cdots \times \mathcal{X}}_{n}$$

and the asymptotic case as:

$$\mathcal{X}^* = \bigcup_{n \geq 0} \mathcal{X}^n. \tag{2.2}$$

[1]The notion of *sample space* is in a way more general than the notions of *alphabet* and *support set*. The connection can be established using the notions of quantization of random variables from Section A.2.3.

Some authors, for example in [Vitányi and Li, 2000], have argued that the MDL principle is meaningful only when related to infinite sequences. The restriction to finite samples does not confer additional theoretical benefits, which is why the following definition uses (2.2).

Definition 2.2 (Probabilistic source)**.** Let $\mathcal{X}$ be a sample space. A *probabilistic source* with values in $\mathcal{X}$ is a function $P : \mathcal{X}^* \to [0, \infty)$ such that for all $x^n \in \mathcal{X}^n$ we have

1. $\sum_{z \in \mathcal{X}} P(x^n, z) = P(x^n)$; and
2. $P(x^0) = 1$

When trying to construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $P = \mathbb{P}$ using Kolmogorov extension theorem, the first condition from Definition 2.2 is called the compatibility condition. The second condition is just a succinct way to express that the total probability should sum to 1.

With all of the above in mind, it is obvious that the probabilistic sources are in fact just distributions over infinite sequences. The terminology introduced above has two benefits:

1. It does not allude to a "true" distribution, which is the case when a probability space is explicitly defined.
2. A probabilistic source can encompass processes with dependence between the elements, like a Markov chains or ARMA-GARCH processes, which are not usually called distributions.

The probabilistic sources can be thought as point elements in a model selection problem. If we have decided that a sample comes from a given probabilistic source, then we have the exact distribution of the sample, hence "know" all of its properties.

It is usually convenient to group probabilistic sources into sets of related sources, and the resulting set is called a *probabilistic model*.

Definition 2.3. A set of *probabilistic sources* is called a *probabilistic model.*

This seems somewhat bare and usually there are more requirements for a probabilistic model in order to be useful. For example the probabilistic sources in a probabilistic model are usually related by similar form and can be indexed by a (vector) parameter θ in some parameter space Θ.

Some examples are given below:

Example 2.4. The following are all examples of probabilistic models. Note that some of them are subset of others.

- The set of all Gaussian distributions with a scale and location parameters.
- The set of all Markov chains with $n = 4$ states.

- The set of all Markov chains with $n \leq 4$ states

Using the above terminology, we make a connection with the standard notions in model selection from statistics:

Definition 2.5 (Point hypothesis)**.** A *point hypothesis* H is the assumption that random variable X is generated by the probabilistic source $P(\cdot|H)$. It is also referred to as a *simple hypothesis*.

Point hypothesis represents the most detailed knowledge we can have about a sample. It is however not always attainable, and for most problems the goal is to find the general form of the distribution of a sample. The following definition expounds this more general concept.

Definition 2.6 ((Statistical) model)**.** A *statistical model* $\mathcal{H}$ refers to a set of probability distributions that have the same general form.

Statistical models are in general probabilistic models from Definition 2.3 with a defined structure, tuned to a particular application. The definitions are necessarily vague in order to encompass the wide variety of models that are used in practice.

Models are usually known as "composite hypothesis". The MDL principle in general does not require a distinction in the two cases above, hence the following definition.

Definition 2.7 (Hypothesis)**.** *Hypothesis* can refer to either a model (Definition 2.6) or a point hypothesis (Definition 2.5).

2.3. Philosophy of the MDL principle

This section displays the basic facts and principles that show how and why the MDL principle is derived from Shannon's information theory, related concepts and general observations on the nature and challenges in statistical inference.

These principles represent identifiable concepts that go beyond the mathematical calculations. As they are intrinsically related, these facts overlap to some extent.

2.3.1. Learning as Data Compression

In the introduction to Section 2.1 an example of the link between patterns (regularity) and compression is presented. This is formalized in the concepts of Kolmogorov complexity (Section A.6), more specifically, Kolmogorov structure function (Definition A.85).

The structure function $K_k(x^n|n)$ (Definition A.85) represents the smallest set that a program with length k can identify that still contains the given sequence.

Recall the first example sequence from Section 2.1:

- 0101010101010101010101010101010101010101...

If the shortest resulting program to describe the program "repeat 500000 times the subsequence xx" has length k is equal to the length, then the structure function

$$K_k(x^n|n) \leq 2$$

because that is the number of bits required to encode the sequence xx=01.

Another way to look at it is that the program "repeat subsequence xx 500000 times" describes a set of four strings:

- 0000000000000000000000000000000000000000...
- 0101010101010101010101010101010101010101...
- 1010101010101010101010101010101010101010...
- 1111111111111111111111111111111111111111...

of which our string is one, and to describe a set of d elements we need $\log_2 d$ bits.

To utilize the structure function, the Kolmogorov minimal sufficient statistic is defined in Definition A.86. The idea is to find which part of the description (program) is essential (captures patterns/regularities) and which is extraneous (captures noise).

In general, by increasing the length of the program by 1 bit we are able to reduce the size of the resulting set in half, which is achieved if that one bit is used to discern in which half of the original set our string was. Since the description of the sequence in the set decreases by 1 bit due to its smaller size, there is no net benefit on the description.

There may be, however, different programs that may produce substantially smaller sets, while having length no more than 1 bit longer than the given program. The length of the program that achieves the shortest total description is called the minimal sufficient statistic, and it is the most compact way to describe the data, i.e. the best model.

The minimal sufficient statistic represents the size of the regularity, and the description length of the remaining set is interpreted as a noise.

A severe drawback of the minimal sufficient statistic and other related Kolmogorov complexity constructs is that it is not computable and can vary wildly for finite strings depending on the computer in question.

For those reasons the MDL principle resorts to a very restricted set of programs, those that represent Shannon-Fano encoded strings and the computer acts as a simple decoder. We connect them with the notions of MDL as follows, with some necessary ambiguity:

- Let $\mathcal{M}_\theta$ be a parametric family described by a parameter $\theta \in \Theta$.
- The probability of the data given the model $f(x^n|\theta)$ (i.e. the likelihood) represents the size of the set that remains to be described once θ is fixed.

- The described set is represented by $f(x^n|\theta)$ - the higher the likelihood, the more likely that x^n belongs to the described set. This is similar to the notion of fuzzy set.
- The minimal sufficient statistic then is the description of $\theta_0 \in \Theta$, for which the length of the description of θ_0 plus the size of the described set that remains is smallest.

2.3.2. Equivalence Between Code and Distribution

Let P be a probabilistic source for the discrete sample space $\mathcal{X}$. From Definition 2.2 it is possible to construct a random variable X with distribution P over $\mathcal{X}$. For each sample $x^n \in \mathcal{X}^n$, the probability of occurrence will be $P(x^n)$.

On the other hand, if we take $C_\aleph$ be a source code with codelength $l(x)$ for X, as defined in Section A.4, then we have the following observations:

1. From Kraft's inequality (Theorem A.53) it follows that
$$\sum_{x^n \in \mathcal{X}^n} D^{-l(x^n)} \leq 1.$$
2. The optimal average codelength L_X is approximately equal to the entropy $H_D(X)$, from Theorem A.57.
3. The optimal average codelength is achieved using Shannon code (described in Example A.61), that has length of codewords approximately equal to the logarithm of the probability:
$$l(x^n) \approx -\log_D P(x^n). \tag{2.3}$$
4. Shannon code is competitively optimal, i.e. it is unlikely to be beaten most of the time, see Theorem A.60.

From all of these points, the following main fact is derived:

Fact 2.8 (Equivalence between code and distribution)**.** *For each probability source (i.e. distribution) there is an optimal code (in expected sense as well as competitively-optimal sense), and it is a Shannon code.*

Conversely, for each source code for the sample space $\mathcal{X}$ there is a distribution $P(x^n)$ satisfying equation (2.3), for which the source code is Shannon code, therefore optimal (in expected as well as competitively-optimal sense, Theorem A.60).

Note that the optimality in Fact 2.8 is somewhat self-referential, i.e. the code is optimal with respect to the distribution only for that distribution. Thus it is no wonder that such a code exists. This equivalence is the great insight in the structure of the relationship between codes and distributions.

2.3.3. Models as Languages

Under the MDL paradigm, the models hypothesized for a sample can be more generally viewed as languages, in that they represent some kind of structure of the sample, allowing (increasing the probability) for some samples or disallowing (decreasing the probability) for others.

For example, if the model H is considered, then the probabilistic interpretation of the codelength $l(x^n|H)$ is that the shorter codewords give exponentially more probability to observe the sequence.

This view also changes the meaning of the word "noise" when attached to a sample. The classical statistical approach to noise is to describe it as a component of the process that has (usually) independent distribution, but is a random variable for all intents and purposes.

According to the MDL philosophy however, the noise is just the bits of description needed to encode the data, after the model H has been used to capture the structure.

Example 2.9 (Markov chain)**.** In a Markov chain, the transition probabilities can be thought of as the structure of the model. Then we can determine a sequence's probability and encode each one with Shannon-Fano code corresponding to it.

Example 2.10 (Regression)**.** If the model chosen is linear regression

$$y = X\beta + \epsilon$$

then the structure of the model is β, and the (hopefully small) regression residual ϵ is the noise.

The best fit is achieved when ϵ has small variance, which corresponds to our ability to encode it to fixed finite precision with fewer bits.

2.3.4. Idealized Codelength

From Section 2.3.2 it is evident that there is a deep link between source codes and discrete distributions. In practice distributions are often continuous, hence the sample space has infinite number of possible values. The connection is established using the quantization of the continuous random variables - for more details see Section 1.1.3.

Thus (2.3) presents an approximate relation between the codelength and the probability of the sample x^n. Equality is only possible when the probabilistic source, viewed as a distribution, is D-adic (see Definition A.55).

For large enough samples, the error in the approximation (2.3) converges to zero, and from Kraft inequality (Theorem A.53) it is not possible to achieve $l(x^n) = -\log_D P(x^n)$ for some x^n, while having the rest of them satisfy $l(x^n) < -\log_D P(x^n)$.

Moreover, what we are interested in the applications of the MDL principle is the codelengths, not the actual codes. Thus we define an idealized codelength and use it instead of a particular exact source coding length.

Definition 2.11 (Idelized codelength)**.** *Idealized codelength* is defined as

$$l(x^n) = -\log_D P(x^n). \tag{2.4}$$

In terms of probabilistic sources (Section 2.2), given that the data x^n follows the probabilistic source attached to the point hypothesis $P(x^n|H)$, the optimal code length will be

$$l(x^n|H) = -\log_D P(x^n|H).$$

It is obvious that (2.4) will not represent a real encoding scheme in general. A real coding scheme will necessarily have some slack to the distribution, because not all codes satisfy Kraft's inequality with equality.

This is not a problem, because the slack can be reduced by encoding consecutive observations together, as a consequence of Corrolary A.58.

2.3.5. Parsimony (Occam's razor)

Another way that the MDL principle can be considered is through Occam's razor, the principle of ontological parsimony. It briefly states that explanations should be kept as simple as possible.

Occam's razor has received much flak on two grounds - that it is arbitrary, and that it is false. Even though the MDL principle is an almost precise embodiment of Occam's razor, it is nevertheless claims less, so is able to give more, in particular with regard to those criticisms

1. **Occam's razor is arbitrary** - since the notion of description is arbitrary on at least two counts - there are many possible languages, after all, and many pre-defined concepts that can dramatically alter the length of the description - then Occam's razor must be arbitrary. That is correct, and the reader may be reminded that this is the major drawback in Kolmogorov minimal sufficient statistic application to model selection as well (Section 2.3.1). That is why the MDL principle uses codes corresponding to probability distributions.
2. **Occam's razor is false** - if we try to model real world situations, they are often very complex, so it is not clear why simple descriptions should be favored. This was put succinctly by G. Webb - "What good are simple models of a complex world?". To answer that, the reader is reminded of the concepts in Section 2.3.3 - models are languages, and description using a model is just that - a description. There is no reason why simple descriptions cannot provide guidance and insight into complex phenomena, after all that is what physical sciences do all the time - simple laws govern complex, even chaotic behavior. Moreover, from a practical side, data is usually limited, so complex models with many details are much harder to deduce, so may not be usable at all.

2.3.6. Data is All We Have

As established in Section 1.3 and Section 1.3.3 the process of model selection, and statistical inference in general, goes beyond mathematical considerations and is often required to incorporate external concerns.

One such concern is Occam's razor, introduced in Section 2.3.5. Another is the epistemological problem - what do we have to presuppose (and consequently, rely on) in order for the statistical procedure to be logically consistent?

The MDL principle presupposes only the existence of the data (the sample). Inference is done using description of the data, and the questions of probability of hypotheses can be re-cast in terms of the length of that description.

Since interpretations of the data require a theory through which the data is viewed, it is a good thing to have a logically consistent procedure that does not depend on the particular theory of the data. For example, if the underlying process is deterministic, using random variables to model is still useful, and can be made internally consistent by interpreting our lack of knowledge as randomness, but that still removes us from the underlying process.

In that sense the MDL procedure is indifferent to the nature of the modeled process - the data is all we have and its description is what we need.

2.4. Crude MDL

The practical MDL, now referred as "Crude MDL" was the historically the first approach to the application of the MDL principle to the problem of model selection.

Let be a list of possible candidate models $\mathcal{H}^{(1)}, \mathcal{H}^{(2)}, \ldots$ to explain data D. The task is to find the model that explains the actually observed data. It is a fairly standard setting for model selection in statistics.

Using the definitions from Section 2.2 we define the following

Definition 2.12 (Crude MDL)**.** The best point hypothesis $H \in \bigcup_k \mathcal{H}^{(k)}$ to explain the data D is the one that minimizes the sum

$$l(H) + l(D|H),$$

where

- $l(H)$ is the description length of the hypothesis; and
- $l(D|H)$ is the description length of the data, when encoded with the knowledge (help) of the hypothesis.

The best model to explain D is then the smallest model containing H.

Usually $\mathcal{H}^{(1)}$ is parameterized by some $\theta^{(k)} \in \Theta^{(k)}$, which depending on the exact specifics may not be hierarchically or otherwise linked for different k.

The above definition is a direct application of the concept of learning as data compression (Section 2.3.1) - the goal is to find the point hypothesis H, that will result in smallest description of the data, while also accounting for the description length of the point hypothesis itself.

The second part is very important and this is in fact what sets apart the MDL principle from simple statistical hypothesis testing. In classical statistics, especially when only two hypothesis are considered, the modus operandi is that all are treated on a similar footing. This however is not always reasonable, in particular when some of the point hypothesis seem very far-fetched - note again the argument from Section 1.1.1.2.

There are two parts of the puzzle that Definition 2.12 is not explicit on:

- how is the data encoded and what is its description length $l(D|H)$; and
- how is the hypothesis described and what is its description length $l(H)$.

Those questions are discussed in the next sections.

2.4.1. Description Length of Data $l(D|H)$

The question of the description length of the data $l(D|H)$ is mostly answered by Shannon's information theory, and in particular the theory of optimal codes.

As was already discussed in Section 2.3.4, given the probabilistic source $P(\cdot|H)$, the optimal codelength is

$$l(\cdot|H) = -\log P(\cdot|H).$$

Other codes are possible, but since they do not guarantee optimality, they are seldom used.

2.4.2. Description Length of Hypothesis $L(H)$

Unlike the description of the data from Section 2.4.1, to which the MDL principle provides reasonable guidelines, the encoding of the hypothesis is where the real conundrum with the "crude" MDL principle lies.

One way is to have some kind of uniformly encoded parameter space. While it is usually useful, it is nevertheless somewhat arbitrary, which is a major critique for the crude MDL.

For example whenever the parameters are also probability of something, as in the case of Bernoulli distribution, the parameter is also the probability of "success".

Thus another sensible choice is to make the log-probability uniformly distributed instead, which gives the parameter interpretation as odds. Such an approach cannot be discarded off-hand, and this ambiguity is problematic.

Nevertheless, due to the description length of the hypothesis in the "crude" MDL principle acts a lot like the penalty term for the other criteria like AIC, BIC, DIC, GIC, etc. The source of that penalty, however, is explicitly derived from the complexity of description of the hypothesis, an idea that culminate as the NML model and model complexity in Section 2.6.

Last but not least, the conceptual simplicity of $L(H)$ allows for some ad-hoc solutions, which may not be proven optimal in theory, but can still provide useful approximations for various real-world problems. Examples from the literature are provided in Section 2.7.1.

2.5. Refined MDL

From the exposition in Section 2.4 it is clear that the crude MDL principle has some problems, mainly by the lack of theoretical restriction on the design for codes for the hypothesis (Section 2.4.2).

To fix that, a natural solution is to design codes that encode both the data and the hypothesis in one go, aptly named one-step codes. To talk about their optimality we first introduce some definitions.

Definition 2.13 (Redundancy)**.** Let P and $\overline{P}$ be probabilistic sources with sample space $\mathcal{X}$. For a given x^n, the *redundancy* (of encoding x^n with $\overline{P}$ instead of P) is defined as

$$RED(\overline{P}, P, x^n) = -\log \overline{P}(x^n) - \{-\log P(x^n)\}\,.$$

An equivalent way to describe Definition 2.13 is the number of extra bits that take to encode x^n using the optimal code for $\overline{P}$ instead of the optimal code for P. With this in mind, while it is possible to have

$$RED(\overline{P}, P, x^n) < 0, \tag{2.5}$$

it is not possible to have (2.5) for all $x^n \in \mathcal{X}^n$, or even on average:

$$\sum_{x^n \in \mathcal{X}^m} RED(\overline{P}, P, x^n) < 0,$$

because it would contradict Kraft's inequality.

Thus we define a measure of difference between the distributions in the following way:

Definition 2.14. Let P and $\overline{P}$ be probabilistic sources with sample space $\mathcal{X}$. The *maximum redundancy* when encoding with $\overline{P}$ instead of P is defined as

$$RED_{\max}(\overline{P}, P) = \max_{x^n \in \mathcal{X}^n} RED(\overline{P}, P, x^n) = \max_{x^n \in \mathcal{X}^n} \left\{ -\log \overline{P}(x^n) - \{-\log P(x^n)\} \right\}.$$

This measure of distance between distributions is used to define universality for distributions:

Definition 2.15 (Universal models)**.** Let $\mathcal{M}$ be a parametric family of probabilistic sources. A sequence of probabilistic sources $\left\{P^{(n)}\right\}$ is called an *universal model* with respect to $\mathcal{M}$, if for any $P \in \mathcal{M}$ and $\epsilon > 0$, there exists $n_0 > 0$ such that $\forall n > n_0$, the maximum redundancy between $P^{(n)}$ and P is bounded on average by ϵ:

$$\frac{1}{n} RED_{\max}(\overline{P}, P) \leq \epsilon \tag{2.6}$$

Each universal model has a corresponding code, defined by the equivalence of code and distribution from Section 2.3.2:

Definition 2.16 (Universal codes)**.** A code corresponding to an universal model (Definition 2.15) is called an *universal code*.

The codelength of universal code U is denoted by $L_U(x^n)$.

This is also referred to as "universal model in individual-sequence sense" in the literature.

Note that n_0 may depend on P and ϵ. This means that there is no limitation with respect to uniformity over $P \in \mathcal{M}$ - for different P and small sample size n the available ϵ may be very large, even unbounded on $P \in \mathcal{M}$.

On the other hand, Definition 2.15 is rather non-restrictive, which will be evident in the next examples of universal models - most of the time we can have even stronger conditions.

The first universal model to be introduced is derived directly from the crude MDL principle (Definition 2.12), the so-called *two-part code*.

Example 2.17 (Two-part code)**.** Let $\mathcal{M}$ be a family of probabilistic sources, indexed by $\theta \in \Theta$. If Θ is finite or countably infinite, then there is a code that encodes the index k of θ_k.

Denote by H_k the hypothesis that x^n comes from the probabilistic source $P(\cdot|\theta_k)$. The length of the code that encodes k is denoted by $l(\theta_k)$.

The two-part code is defined to have codelength

$$L_{2-p}(x^n) = \min_k \{-\log P(x^n|\theta_k) + l(\theta_k)\}$$

and its equivalent universal model can be derived from the equivalence between code and distribution (2.3):

$$P_{2-p}(x^n) = \exp\left(-\min_k \{l(\theta_k) - \log P(x^n|\theta_k\}\right) = \max_k \{P(x^n|\theta_k)\exp(-l(\theta_k))\}.$$

Then (2.6) is satisfied directly:

$$\begin{aligned} RED(P_{2-p}^{(n)}, P_{\theta_k}) = & \min_k \{l(\theta_k) - \log P(x^n|\theta_k)\} + \log P(x^n|\theta_k) \\ \leq & \; l(\theta_k) - \log P(x^n|\theta_k) + \log P(x^n|\theta_k) \\ = & \; l(\theta_k) \end{aligned}$$

In order to define a two-part code for uncountably infinite Θ, it is necessary to first partition it as $\{\Theta_k\}$ in such a way that choosing for each k a single $\theta_k \in \Theta$ will ensure that $RED(P, P_{\theta_k})$ will be very small. Then declare that the only possible values for θ are $\{\theta_k\}$ and proceed with the above construction.

The next example of a universal models directly applies the concept of Bayesian inference from Section 1.1.1.3.

Example 2.18 (Bayesian universal model)**.** Let $\mathcal{M}$ be a family of probabilistic sources, indexed by $\theta \in \Theta$. Define an arbitrary probability distribution $\pi(\theta)$ with the only requirement that $\pi(\theta) > 0, \forall\theta \in \Theta$, that will be interpreted as the prior probability of the parameters (see Section 1.1.1.3).

Then the Bayesian universal model $P_{Bayes}^{(n)}$ is defined as

$$P_{Bayes}^{(n)} = \sum_{\theta\in\Theta} P(x^n|\theta)\pi(\theta).$$

Because

$$\begin{aligned} RED(P_{Bayes}^{(n)}, P_{\theta_0}) = & -\log P_{Bayes}^{(n)}(x^n) - \{-\log P_{\theta_0}(x^n)\} \\ = & -\log\left[\sum_{\theta\in\Theta} P(x^n|\theta)\pi(\theta)\right] - \{-\log P_{\theta_0}(x^n)\} \\ \leq & -\log\left[P(x^n|\theta_0)\pi(\theta_0)\right] - \{-\log P_{\theta_0}(x^n)\} \qquad (2.7) \\ = & -\log\pi(\theta) \qquad (2.8) \end{aligned}$$

It follows that the redundancy is limited by a constant dependent only on θ, hence the Bayesian universal model is universal.

Moreover, in case of a finite Θ, there exists $\pi_0 = \min_{\theta\in\Theta}\pi(\theta) > 0$, which bound all redundancies via (2.8), hence even a stronger property than (2.6) is observed - uniform bound over $P \in \mathcal{M}$!

On the other hand, for infinite Θ there is no such π_0, so we cannot find an uniform bound, but still Definition 2.15 is in effect, because for each θ we can show that

$$n_0 = \left\lceil \frac{-\log \pi(\theta)}{\epsilon} \right\rceil$$

satisfies (2.6).

This example exemplifies to the main limitation of universal models - when the model $\mathcal{M}$ has too diverse distributions, i.e. ones that encode a sample x^n with very different codelengths, then the redundancy can vary wildly and it is a real challenge to find a universal model that will keep the redundancy low enough to be practical.

The particular problem with the Bayesian model is that for distributions that we have assigned a very low prior probability the redundancy limit (2.6) can be particularly useless, since $-\log \pi(\theta)$ is not limited from above.

On the other hand, the inequality (2.7) is not the best bound that can be found, especially when the distributions are similar. It may be even desirable to reduce an uncountably infinite set to a finite set by restricting the prior probabilities to be nonzero only on a finite grid, and then use that prior for basis in a Bayesian universal model.

Finally, why study the Bayesian universal model? First of all, it is better than the two-part:

Lemma 2.19. *Let $\mathcal{M}$ be a finite family of probabilistic sources, indexed by $\theta \in \Theta$. Then for every two-part universal model L_{2-p} there exists a Bayesian universal model L_{Bayes} with uniformly shorter codelengths:*

$$\forall x^n \in \mathcal{X}^n : L_{2-p}(x^n) \geq L_{Bayes}. \tag{2.9}$$

The inequality (2.9) is necessarily strict for some x^n whenever $\#\{\Theta\} > 0$.

Proof. Let $l(k)$ be a code encoding the hypothesis $\theta = \theta_k$. In this case there is a prior probability

$$\pi(k) = e^{-l(k)}$$

and we can define

$$P_{Bayes} = \sum_k P(x^n|\theta_k) e^{-l(k)}.$$

Let x^n be a sample and k_0 achieves $\min_k \{l(H_k) - \log P(x^n|H_k)\}$. Then

$$\begin{aligned} L_{2-p}(x^n) &= l(k_0) - \log P(x^n|H_{k_0}) \\ &= -\log P(x^n|H_{k_0})\pi(k_0) \\ &\geq -\log \left[\sum_{\theta \in \Theta} P(x^n|\theta)\pi(\theta) \right] \\ &= P_{Bayes}. \end{aligned}$$

The inequality again follows from the fact that the sum is larger than any of its parts. □

With the help of universal models then we can transform Definition 2.12 into the following General Refined MDL Principle for model selection.

Definition 2.20 (Refined MDL)**.** The best model from a set of models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ is the one that minimizes the description length

$$L(x^n|\mathcal{M}_i),$$

where the codelength is calculated using a pre-specified universal model.

2.6. Model Complexity and Normalized Maximum Likelihood

This section describes in more details the Normalized Maximum Likelihood (NML) universal model, first derived in [Shtarkov, 1987].

The exposition starts from measures of closeness between distributions ("redundancy"), proceeds to define the NML and its model selection application - the Stochastic Complexity(SC) criterion.

In conclusion the challenges of calculating model complexity are explored with the purpose of introducing the motivation for the distribution complexity in Chapter 3.

2.6.1. Regret

The redundancy metric from Definition 2.13 quantifies the difference between two distributions, as far as a single sample x^n is concerned. To extend that over $\forall x^n \in \mathcal{X}^n$ the maximum redundancy was defined (Definition 2.14), and the concept of universal models was introduced.

The rest of the discussion and particularly the Bayesian universal model from Example 2.18 pointed out that for universal models in general we do not have guarantees of the redundancy - for some distributions in the model the redundancy between the distribution and the universal model can get arbitrary high, or equivalently, the loss suffered by using a universal model, instead of the best-fitting model for the sample in question, is not bounded.

To quantify that concept we utilize a min/max approach, called the *regret*. First on the minimization part:

Definition 2.21 (Regret)**.** Let $\mathcal{M}$ be a class of probabilistic sources. Let $\overline{P}$ be a probabilistic source sample space $\mathcal{X}$, not necessarily in $\mathcal{M}$. For a given x^n, the regret of $\overline{P}$ relative to $\mathcal{M}$ is defined as

$$\mathcal{R}^{\mathcal{M}}(\overline{P}, x^n) = \min_{P \in \mathcal{M}} RED(\overline{P}, P, x^n) = -\log \overline{P}(x^n) - \min_{P \in \mathcal{M}} \{-\log P(x^n)\}$$

An equivalent way to describe Definition 2.21 is the number of extra bits needed to encode x^n using $\overline{P}$ instead of the optimal (best-fitting) *with hindsight* distribution from $\mathcal{M}$.[2]

This again is specific for each x^n, i.e. different samples will have different $P \in \mathcal{M}$ taken when computing the regret.

Moreover, while it is possible to have $\mathcal{R}^{\mathcal{M}}(\overline{P}, x^n) < 0$, it is not possible to have it for all $x^n \in \mathcal{X}^n$ or even $\sum_{x^n \in \mathcal{X}^m} \mathcal{R}^{\mathcal{M}}(\overline{P}, x^n) < 0$, because it would contradict Kraft's inequality. This is so because $\min_{P \in \mathcal{M}} \{-\log P(x^n)\}$ is not an achievable code - it is not uniquely decodable, since in order to decode it, we must know which sequence is used to determine the optimal code for that sequence.

On the other hand for some sample x^n it is possible that the regret $\mathcal{R}^{\mathcal{M}}(\overline{P}, x^n)$ is very high, and one of the basics of the MDL principle is that the data is all we have (Section 2.3.6). To guard against such possibility, the regret defined in Definition 2.21 should be aggregated over x^n and restricted.

There are a lot of ways to define aggregations for the "optimality" of $\overline{P}$, and one is the *minimax regret*:

Definition 2.22 (Maximum (worst-case) regret)**.** The worst-case regret for the distribution $\overline{P}$ with respect for the probabilistic model $\mathcal{M}$ is defined as

$$\mathcal{R}^{\mathcal{M}}_{\max}(\overline{P}) = \max_{x^n \in \mathcal{X}^n} \left\{ \mathcal{R}^{\mathcal{M}}(\overline{P}, x^n) \right\}.$$

It is also called *minimax* regret.

The maximum regret is a measure of how well the distribution $\overline{P}$ approximates all the distributions in $\mathcal{M}$. Since it considers all distributions in the model $\mathcal{M}$ in an uniform way, a very large regret may be either because the distribution $\overline{P}$ is very distant from the models in $\mathcal{M}$, or because the models in $\mathcal{M}$ are very diverse.

A well-constructed universal model $\overline{P}$ must allow regret to increase only due to the diversity of $\mathcal{M}$.

There are other ways to aggregate the regret that considers all possible samples x^n, for example by finding the expected (or average) regret by supplying $\mathcal{X}^n$ with a probability measure, usually uniform. Limiting expected regret, however, does not guarantee that the regret for a particular sequence x^n will be small, so it is less useful.

[2] It is not in possible to use this best-fitting distribution to encode the sample, because it is only evident which is the optimal *after* the sample is known. Still, the difference provides a uniform criterion for the closeness of the universal distribution to all distribution in $\mathcal{M}$.

2.6.2. Normalized Maximum Likelihood

One particularly useful universal model is Shtarkov's NML. It was introduced in [Shtarkov, 1987] and is also based on the likelihood principle (Section 1.1.1.1).

Definition 2.23 (Shtarkov's Normalized Maximum Likelihood (NML))**.** Let $\mathcal{M}$ be a family of distributions, indexed by a (vector) parameter $\theta \in \Theta$, whose MLE $\hat{\theta}$ is well-defined and unique for all data samples x^n.

Let the following integral be finite:

$$\int_{y^n \in \mathcal{X}^n} P(y^n|\hat{\theta}(y^n))dy^n < \infty.$$

Then the model defined by normalizing the likelihood $P(x^n|\hat{\theta}(y^n))$:

$$P_{NML}(x^n) = \frac{P(x^n|\hat{\theta}(x^n))}{\int_{y^n \in \mathcal{X}^n} P(y^n|\hat{\theta}(y^n))dy^n}$$

is called the *Normalized Maximum Likelihood (NML)* model.

The above definition only mentions absolutely continuous distributions, but with procedure like quantization, it can be applied to discrete distributions as well. In this case, instead of the p.d.f., P is taken to be the p.m.f.

Lemma 2.24 (Properties of the NML)**.** *$P_{NML}(x^n)$ satisfies the following properties:*

1. *P_{NML} achieves constant regret for any sample x^n, i.e. $\mathcal{R}^{\mathcal{M}}(P_{NML}, x^n) = const$; hence*
2. *$P_{NML}(x^n)$ is a universal model.*

Proof. Since $\hat{\theta}$ is the MLE, then from Definition 1.16 it is the solution of the equivalent expression

$$\hat{\theta} = \arg_\theta \min \{-\log P(x^n|\theta)\}.$$

From the definition of regret and the NML distribution:

$$\begin{aligned}
\mathcal{R}^{\mathcal{M}}(P_{NML}, x^n) = & \quad -\log P_{NML}(x^n) - \min_{P \in \mathcal{M}} \{-\log P(x^n)\} \\
= & \quad -\log P_{NML}(x^n) - (-\log P_{\hat{\theta}}(x^n)) \\
= & \quad \log \int_{y^n \in \mathcal{X}^n} P(y^n|\hat{\theta}(y^n))dy^n.
\end{aligned}$$

Therefore the regret is constant.

Since the redundancy is smaller than the regret, then it is bounded by this constant, so the NML is a universal model according to Definition 2.15. □

The constant regret realized by the NML model is also known as *model complexity*.

Definition 2.25 (Model complexity)**.** Let P_{NML} exist for a parametric family $\mathcal{M}$. Then the model complexity of $\mathcal{M}$ is defined as

$$COMP(\mathcal{M}) = \log \int_{y^n \in \mathcal{X}^n} P(y^n|\hat{\theta}(y^n))dy^n.$$

The model complexity has some very interesting properties, in particular when $\mathcal{M}$ is finite, $COMP(\mathcal{M})$ is equal to the logarithm of the number of "essentially different" distributions in $\mathcal{M}$. Some authors call this *counting interpretation*, and can be seen in the following way.

Let $\mathcal{M}$ be a finite family parameterized by $\theta \in \Theta$. Then

$$\begin{aligned}
\exp COMP(\mathcal{M}) &= \sum_{y^n \in \mathcal{X}^n} P(y^n|\hat{\theta}(y^n)) \\
&= \sum_{\theta \in \Theta} \sum_{y^n \in \mathcal{X}^n, \hat{\theta}(y^n)=\theta} P(y^n|\theta) \\
&= \sum_{\theta \in \Theta} \left\{ 1 - \sum_{y^n \in \mathcal{X}^n, \hat{\theta}(y^n)\neq\theta} P(y^n|\theta) \right\} \\
&= |\Theta| - \sum_{\theta \in \Theta} \sum_{y^n \in \mathcal{X}^n, \hat{\theta}(y^n)\neq\theta} P(y^n|\theta)
\end{aligned}$$

The last term is the sum of the probabilities that one distribution will be mistaken for another.

If two distributions, say those derived by θ_i and θ_j, give essentially the same likelihood, the probability of mistaken identification between them will be close to 1. This means that the complexity of the model will not be affected.

This also means that complexity is not a complete guide if model identification is also required of the model selection procedure (see Section 1.3.3).

The problem is usually sidestepped when the parametric family is constructed such that the internal logic of the parameterization provides that for large enough samples, the distributions will be distinguishable.

There are other interpretations of the model complexity:

- *Bayesian interpretation* - NML coincides with Bayesian model if Jeffreys prior is used.
- *Compression interpretation* - NML is a good summary of $\mathcal{M}$ - the better the best model in $\mathcal{M}$ fits the data, the shorter code P_{nml} provides, and P_{nml} treats all models in $\mathcal{M}$ on equal footing.
- *Prequential interpretation* - MDL gives preference to models that compress sequentially produced data in an optimal way, like NML.

More information and further examples of the four interpretations can be found in [Grunwald, 2004, Grünwald et al., 2005, Grunwald, 2007].

The last property we consider here is that when $COMP(\mathcal{M}) < \infty$ and under some mild regularity conditions on the model $\mathcal{M}$, the following asymptotic expansion is in effect:

$$COMP(\mathcal{M}) = \frac{k}{2}\log\frac{n}{2\pi} + \log\int_{\theta\in\Theta}\sqrt{|I(\theta)|}d\theta + o(1),$$

where $I(\theta)$ is the Fisher information matrix.

The residual term $o(1)$ is equal to zero for exponential families, but in general can be arbitrarily high, so the above formula is not always useful for small samples.

2.6.3. Stochastic Complexity Criterion

As developed in [Rissannen, 1989], the Stochastic Complexity criterion (referred as SC later in the text) is broadly speaking the use of model complexity for statistical inference.

It is a direct application of the Crude MDL principle from Definition 2.12, but considering a countable or finite set of possible models. It can be expressed as follows:

Definition 2.26 (Stochastic Complexity criterion)**.** Let $\mathcal{M}_1, \mathcal{M}_2, \ldots$ be a set of possible (universal) models for a sample x^n. The description length of x^n for each model can be defined as

$$SC(\mathcal{M}_i) = -\log P_{\mathcal{M}_i}(y^n|\hat{\theta}(y^n)) + COMP(\mathcal{M}_i).$$

The best model is the one with minimum $SC(\mathcal{M}_i)$.

This also parallels the definitions of AIC from Example 1.33 (and the other similar criteria) in that we calculate a number $SC(\mathcal{M}_i)$ that quantifies the how good the particular model is in terms of both goodness of fit (log-likelihood) and generalizability (model complexity).

Another way to view Definition 2.26 is as direct consequence of using an NML universal model in the Refined MDL principle from Definition 2.20.

The next section will discuss the main challenge in the application of Definition 2.26, and the main contribution of this dissertation will a new solution for the case of scale-location families, which is described in detail in Chapter 3.

2.6.4. The Infinity Problem (a.k.a. Singularity)

The main challenge for a lot of problems is the fact that the model complexity from Definition 2.25 is infinite. This is often called the *infinity problem*, and it is an obstacle to the use of model complexity in statistical inference.

There are several solutions, which arise partly from different interpretations of the reasons for the infinity problem. They are discussed in the following sections.

First we start by introducing variants of the model complexity, namely the constrained model complexity.

Definition 2.27 (Constrained complexity, by sample space)**.** Let $\mathcal{A} \subseteq \mathcal{X}^n$ and $\mathcal{M}_\theta$ be a parametric family. The constrained model complexity of restricting $x^n \in \mathcal{A}$ is defined as

$$COMP(\mathcal{M}|x^n \in \mathcal{A}) = \log \int_{y^n \in \mathcal{A}} P(y^n|\theta = \hat{\theta}(y^n))dy^n. \tag{2.10}$$

There is another way to construct a constraint on the complexity, and that is by limiting the values that parameter θ can achieve. Some authors, like in [Grunwald, 2007] do not call this a constraint and just bundle it with the usual definition of the model complexity, because in the end the possible values of θ determine $\mathcal{M}_\theta$.

This type of restriction will be considered in this section only, and we will refer to it by the following definition:

Definition 2.28 (Model complexity, restricted by parameter space)**.** Let $\hat{\theta}$ be a ML estimator for a parametric family $\mathcal{M}_\theta$ and let $\Theta_0 \subseteq \Theta$. Define

$$\hat{\theta}^* = \arg_{\theta \in \Theta_0} \max P(y^n|\theta).$$

Then the model complexity restricted by parameter space of $\mathcal{M}$ is defined as

$$COMP(\mathcal{M}|\theta \in \Theta_0) = \log \int_{y^n} P(y^n|\theta = \hat{\theta}^*(y^n))dy^n.$$

Note that because $\hat{\theta} = \hat{\theta}^*$ for all x^n for which $\hat{\theta} \in \Theta_0$, the difference in model complexity comes from the boundary effect of snapping the estimate to the boundary of Θ_0.

Both restriction types lead to similar behavior, but have very different philosophical implications. We will look at their differences now.

Let Θ_0 be a fixed set and define $\mathcal{A} = \left\{x^n : \hat{\theta}(x^n) \in \Theta_0\right\}$. In this case

$$\begin{aligned}\exp COMP(\mathcal{M}|\theta \in \Theta_0) = & \int_{y^n} P(y^n|\theta = \hat{\theta}^*(y^n))dy^n \\ = & \int_{y^n \in \mathcal{A}} P(y^n|\theta = \hat{\theta}(y^n))dy^n \\ & + \int_{y^n \notin \mathcal{A}} P(y^n|\theta = \hat{\theta}^*(y^n))dy^n \\ = & \exp COMP(\mathcal{M}|x^n \in \mathcal{A}) \\ & + \int_{y^n \notin \mathcal{A}} P(y^n|\theta = \hat{\theta}^*(y^n))dy^n.\end{aligned}$$

This means that the model complexity when restricting parameter space is equal to the suitably chosen restriction of the sample space, plus a multiplicative term.

Now we introduce the Gaussian model, which serves as an example.

Example 2.29 (Gaussian model, fixed variance)**.** Let $\mathcal{M}$ be a Gaussian parametric family with known σ_0, i.e. $P(x^n|\mu) \sim \mathcal{N}(\mu, \sigma_0)$. In this case the integral

$$COMP(\mathcal{M}) = \log \int_{y^n \in \mathcal{X}^n} P(y^n|\hat{\mu}(y^n))dy^n = \infty.$$

If we restrict the integration to a subset of y^n, for example $\hat{\mu} \in [a; b]$, then the constrained complexity by sample space will be finite (see for example [Barron et al., 1998, Stine and Foster, 2001]):

$$COMP(\mathcal{M}|\hat{\mu} \in [a; b]) = \log \int_{\hat{\mu}(y^n) \in [a;b]} P(y^n|\hat{\mu}(y^n))dy^n = \log \left(\frac{b-a}{\sqrt{2\pi}\sigma} \cdot \sqrt{n} \right). \quad (2.11)$$

This provides an NML-type model:

$$P^K_{NML}(x^n) = \frac{P(x^n|\hat{\mu}(x^n))}{\int_{\hat{\mu}(y^n) \in [a;b]} P(y^n|\hat{\mu}(y^n))dy^n}$$

This model however depends on the boundaries a and b in significant ways, so it is not possible to use the resulting codelength to compare models.

The following subsections consider ways in which to remove the dependence on the boundaries.

2.6.4.1. Renormalization

The dependence on the boundaries necessitates some form of extension of the code to account for that, which some authors refer to as meta-universal model (see [Grunwald, 2004]).

This two-part coding idea was pioneered in [Rissanen, 1999]:

1. Use constrained complexity, limiting the parameters in some "natural" way, for Example 2.29 use $\hat{\mu} \in [-K; K]$.
2. Use K as a hyper-parameter, that is, using a two-part code first describe the smallest integer K that is sufficient for the data at hand, using arithmetic coding for example.
3. With that K fixed, utilize a constrained NML model.

The resulting code will have codelength of the form

$$L_{RNML}(x^n) = -\log P_{RNML}(x^n) = L(K(x^n)) - \log P^K_{CNML}(x^n)$$

This code is called Re-Normalized Maximum Likelihood ([Rissanen, 2000]).

The advantage of this method is that it is relatively easy to compare two models, because what we receive in the end is a complete code for a sample. It also corresponds to a universal modell.

A distinct disadvantage of this procedure is that it is lacking region indifference (i.e. equal treatment of models from $\mathcal{M}$), which was basis for the compression interpretation in Section 2.6.2. This is not an insurmountable problem, but somewhat negatively reflects on the elegance of the notion of model complexity.

2.6.4.2. Restricting parameter space

Another idea, introduced in [Stine and Foster, 2001] is to restrict the parameter space by introducing a different estimator $\hat{\theta}_{\Theta_0}$, which is still MLE but reduces Θ to a somewhat more manageable Θ_0, depending on the current problem.

For the Gaussian model (Example 2.29) the effect is to define the estimator as

$$\hat{\mu}_{[a;b]} = \arg_\mu \max_{a \le \mu \le b} f(x^n|\mu),$$

and the model complexity becomes

$$COMP(\mathcal{M}|a \le \mu \le b) = \log \int_{y^n \in \mathcal{X}^n} P(y^n|\hat{\mu}_{[a;b]}(y^n))dy^n$$

The effect is to add another term to (2.11):

$$\begin{aligned} \exp COMP(\mathcal{M}|a \le \mu \le b) = & \int_{y^n \in \mathcal{X}^n} P(y^n|\hat{\mu}_{[a;b]}(y^n))dy^n \\ = & \int_{\{y^n \in \mathcal{X}^n\} \cap \{\hat{\mu}(y^n) \in [a;b]\}} P(y^n|\hat{\mu}(y^n))dy^n, \quad (2.12) \\ & + I(\{\hat{\mu}(y^n) < a\}) + I(\{\hat{\mu}(y^n) > a\}) \quad (2.13) \end{aligned}$$

where the parameter of I is a set, and

$$I(S) = \int_{\{y^n \in \mathcal{X}^n\} \cap S} P(y^n | \partial S) dy^n.$$

It is easily seen that $I(S)$ is finite, because the function under the integral is a probability, and the integral represents the probability that for a sample y^n the estimate of $\hat{\mu}_{[a;b]}$ lies on the boundary, i.e. the unrestricted estimator $\hat{\mu}$ is beyond a or b.

For multidimensional parameter vectors θ the relation (2.12) obtains a whole new level of complexity, as for some data samples only one parameter is effectively restricted, and a lot more combinations have to be taken into account.

The disadvantage of restricting the parameter space is that still a and b will have to be selected in some way. If we have theoretical reasons to presuppose one interval $[a; b]$ over another, then we can use this method. If we try to code around this limitation like in Section 2.6.4.1, we still obtain preference of one code over another.

Moreover, the parameters θ usually do not correspond from one model to another, for example, in a Student-T distribution the σ is not comparable to a Gaussian distribution σ, although both are location paramerters. This is accounted for in the distribution complexity in the next section.

2.6.4.3. Distribution complexity

A third way to effectively solve the infinity problem is the distribution complexity. This is the main result in this dissertation and is presented in greater details in Chapter 3.

In some cases, there are closed-form solutions for the distribution complexity. For cases where such is not possible, Chapter 4 discusses how numerical calculations can be performed to find a solution.

The basic idea is that we still want region indifference like in Section 2.6.4.2, and the idea of restricting the sample space Y^n is more intuitive than restricting the parameter space like in Section 2.6.4.1.

The idea is to find how the hyper-parameters K match between models, that is, how to make the boundaries of the sample space Y^n the same between the models. Then we can dispense with the hyper-parameters directly, and do inference with the differences in complexity.

This reminder term, obtained after matching the contributions of the boundaries, is called the *distribution complexity* and is introduced in Chapter 3.

2.7. Recent Advances in MDL

This section lists systematically the recent advances in the field in terms of theory, practice and philosophy of the MDL Principle.

We start in Section 2.7.1 with a focus on the direct applications of the MDL principle. The examples given revolve around solutions of model selection problems, often by analyzing *big data*, and a diverse selection of examples is provided. It is focused on the properties exhibited by MDL-based procedures in practice.

There are also related applications of similar concepts in articles that do not mention MDL principle explicitly, but adhere to its same philosophy, for example the application of Variable-Length Markov Chain (VLMC) for statistical inference.

In Section 2.7.2 the attention is switched to the theoretical advances in the field of MDL. Although the articles and concepts described in the cited articles are no less practically useful than those in Section 2.7.1, their characterizing property is the wide implications they have for the field and the elegance of the concepts discussed.

Last in Section 2.7.3 are discussed some recent advances in the philosophical ideas behind the MDL principle. The strong connections between the MDL philosophy and Kolmogorov complexity are reaffirmed, and some critical conclusions are drawn.

We finish with a brief recapitulation in Section 2.7.4.

2.7.1. Practice

In this section are presented some recent direct applications of the MDL principle.

The article [Ramírez and Sapiro, 2012b] describes an MDL treatment of the low-rank data modeling problem - how to find the "true" low-rank data matrix, which may be buried under a lot of noise. Since noise is not always Gaussian, it can severely impact classical principal component analysis (PCA), so the authors' use Robust PCA (see [Candès et al., 2011] for more information). This point is important because adherence to the MDL philosophy necessitates the consideration of real-world phenomena, where deviations from normality are the norm.

Another useful property of the MDL principle that [Ramírez and Sapiro, 2012b] exploits is the automatic way in which the inference can be performed, with the authors noting that classical methods for low-rank data modeling require the fine-tuning of several parameters, for which there is no objective criterion. MDL provides such a criterion - the encoded data length minimization.

Let Y be the observed data matrix. The assumption is that

$$Y = X + E,$$

where X is the "real" underlying matrix X.

The authors are explicit in their utilization of two-part codes. They design an explicit direct coding scheme for the rank of X. Because of its low rank, X is decomposed by singular value decomposition (SVD), and then each part of the decomposition is encoded separately. Coding for the residual E is also provided, with the consideration that it may exhibit sparsity.

Another interesting point to take is that the codes designed should take into account the smoothness of some variables/components, which is important for images.

The case study in the paper is performed Robust PCA on moving images with the intent to derive several main components that turn out to be the background and lighting of the scene, and the noise remains the moving figures of people. The smoothness of the movement of the people is also taken into account in the coding scheme.

Although the codes are very well designed, a drawback of the two-part code scheme is that it is not as good as a Bayesian universal model, if assigning a prior is computationally feasible. Another concern is that even with the great lengths at which the authors provide justification for the given coding scheme, it can still be deemed somewhat arbitrary.

A more general frameworks for learning in a sparse coding environment with an MDL-based approach is presented in [Ramírez and Sapiro, 2012a]. Similarly to [Ramírez and Sapiro, 2012b] two-part codes are used and explicit coding schemes are designed. This is not usually done in an MDL settings, because only the code-lengths are useful for statistical inference. The authors also take into account the quantization effects at every step of the encoding.

Additionally, in contrast to [Ramírez and Sapiro, 2012b], the robustness is now incorporated via the l_1 norm.

As a lot of work has been done on sparse models previously, authors present justification why the MDL procedure proposed is better in this case - because in the traditional sparse modeling procedure there are a lot of parameters to be tuned. When tuning parameters it is best to have intuitive justification why the parameters should be thus. MDL provides this via the notion of knowledge = compression.

The problem of sparse dictionary estimation can be defined as

$$\mathbf{y} = \mathbf{D}\mathbf{a} + \mathbf{e}, \mathbf{D} \in \mathbb{R}^{m \times p},$$

where $\mathbf{D}$ is the dictionary, $\mathbf{a}$ is the fitted coefficients and $\mathbf{e}$ are the residuals, commonly interpreted as noise. Then the code for the data is defined as

$$l(\mathbf{y}, \mathbf{a}, \mathbf{D}) = l(\mathbf{e}|\mathbf{a}, \mathbf{D}) + l(\mathbf{a}|\mathbf{D}) + l(\mathbf{D}),$$

and each of the codes is explicitly designed, quantization effects included. For some of codes a Bayesian mixtures of exponents is used, which is more efficient.

The authors define three algorithms for MDL-based learning - a Codelength Minimizing Pursuit Algorithm (COMPA) and two algorithms for dictionary learning - one with fixed size p and the other for unspecified size. There is discussion on the computational complexity (hence, feasibility in practice) of the algorithms.

The authors discuss four case studies, all representing problems in image recognition, one of which was already presented in [Ramírez and Sapiro, 2012b].

The overall contribution of the paper is the general MDL framework for sparse codes, with incorporated effects of the quantization of continuous random variables. The case studies may need to be expanded with a more diverse set of applications.

Another interesting topic with direct application is the fitting of linear mixed models, introduced in [Liski and Liski, 2008b]. The authors provide a basic framework for a normalized maximum likelihood (NML) code, to which they refer to as "two-part code". The NML code is conditioned on some of the parameters. This is a standard procedure and the parameters in question can be considered a variety of hyper-parameters.

The authors compare several procedures for model selection for linear mixed models with several classical criteria - AIC, BIC, MDL - and claim that the performance of the MDL-derived criterion is close to that of BIC. They also mention that they have simulation results, but do not present them here. They are described in [Liski and Liski, 2008a], along with application to spline modeling.

These simulation results showing significant performance gains in mean-squared error when model selection is performed via their MDL-based procedure. Still, one would want to see more details on the results and comparison of the model selection criteria, as well as application to real experimental data.

In [Corcoran et al., 2014] another improvement in the application of an MDL-based learning is described, in this case to estimate the parameters and structure of Bayesian networks. Learning those is solvable for discrete or Gaussian data in a variety of ways, but for non-Gaussian continuous data is much harder, and the key is good discretization of the variables.

In order to compute optimal discretization, an MDL-based method was originally proposed in [Friedman and Goldszmidt, 1996], but it suffered from very high computational costs.

The improvement in [Corcoran et al., 2014] is to find a faster algorithm, making the application of the MDL-based method of [Friedman and Goldszmidt, 1996] computationally feasible.

The codes used in the computation are explicit two-part codes. The best code for each step of the encoding is separately developed.

Simulated data is used to showcase that the learning algorithm correctly identifies the Bayesian network used to generate the data, but it would be preferable have an example involving real data. Particular concern is the size of the simulated sample - 100 000 elements, which may not be available for a lot of practical inference problems.

The next cluster of connected papers ([Malyutov et al., 2013a], [Ryabko et al., 2010] and [Malyutov et al., 2013b]) describe a Variable-Length Markov Chain (VLMC) method to test for homogeneity of time-series.

VLMC represent stochastic processes that have different length of memory, that is depending on the current state, the history that influences the next state is of different length, whereas Markov chains have fixed length.

VLMC is also a method for efficient coding, introduced in [Rissanen, 1983], and is usually defined in terms of context trees. The size of the tree is dynamically adapted to correspond to the given sample, which is where the minimum length of the description connects it to the MDL basic principle.

While not mentioning MDL by name, the underlying basic principles outlined in Chapter 2 are observed.

The basic idea of [Malyutov et al., 2013b] is that if the compressed length from a VLMC-encoded data sample is similar to that of the training sample, then the two have similar behavior.

Two examples are developed - one trying to find volatility clustering in NASDAQ data, and the other to determine seismic events using helium emissions. In both cases the usage of compression leads to clear identification of the points where homogeneity of the time-series breaks, which is where the important events in both example cases happen.

A somewhat more general framework on the application of compression-based methods in time-series analysis is described in [Ryabko et al., 2010].

[Ridder and Pintelon, 2005] tries to work through some tough problems with overfitting for very small sample sizes.

The authors make analogy with the Akaike's information criterion (AIC) and more importantly its small sample corrections - AIC_c and AIC_u. In AIC the goodness of fit is independent of the number of parameters and the penalty is proportional to the number of parameters, in AIC_c the penalty function is strongly sloping upwards for small sample sizes, and in AIC_u the goodness of fit is decreasing function of number of parameters.

The justification of those modified criteria is classically done by considering second-order approximations during derivation (whereas AIC considers only linearity), as in [Bedrick and Tsai, 1994], or by using unbiased estimate for Kullback-Leibler (K-L) divergence, as in [McQuarrie et al., 1997].

By the same considerations the authors derive a modified MDL-based criterion is derived, called MDL_c for which

$$MDL_c : -\ln f(x^n|\hat{\theta}) - \frac{(N \log N)\,(p+1)}{N-p-2}$$

where N is the sample size and p is the number of free parameters. Additional criteria AIC_s and MDL_s are provided when the residual variance is known in advance.

The authors provide simulation results that show that in certain cases MDL_c outperforms the other criteria. Further investigation on the theoretical justification may be desired - MDL and AIC are based on different ideas, so it is not directly evident why corrections to AIC should translate to equivalent corrections for an MDL-based criterion.

[de Brauwere et al., 2005] introduces an alternative to information criteria like MDL and AIC when doing model selection for regression where the residuals exhibit heteroscedasticity, and the exact variance of the residuals is known.

The natural choice for the parameter estimation in this case is weighted least squares (WLS). The difference from MDL is in how the residual variance is ascertained to be adequate. Because their variance is known, it is easy to create a χ^2 distributed statistic, and calculate a corresponding p-value. It is then compared to a prespecified confidence level to filter out all underspecified models.

The authors themselves admit to the disadvantages of their model selection routine - a lot of information has to be known in advance, which is usually derived from the model in question - the noise variance - and it can vary between residuals.

Moreover, it may be entirely inconsequential to know the noise variance in a model selection setting, especially if the model is misspecified - a common case in real applications.

An advantage of the method is that the criterion quantifies the absolute merits of the model, rather than the relative ones obtained from MDL and AIC. The cost for that is the requirement that the user provides a confidence level against which the p-value is compared to make sure that a given model is not underspecified.

Another topic where the MDL and other entropy-based procedures are under investigation is signal detection and source number enumeration. Recent article in the topic include [Asadi and Seyfe, 2013], [Haddadi et al., 2010] and [Nadler, 2010].

[Haddadi et al., 2010] improves over the classical work [Fishler et al., 2002] by providing a more accurate characterization of the probability of underspecification, i.e. missed detection of a signal. Simulation results show that the proposed analysis is better than previously available approximations.

Further progress was done by [Nadler, 2010], where a statistical performance analysis of MDL and AIC-based estimators of the number of signal sources was done. A modified AIC-type estimator which outperforms the MDL estimator is proposed, which unlike in [Haddadi et al., 2010] has an explicit formula for the correction term.

The results in [Haddadi et al., 2010] are again based on simulations. What remains to be seen is if similar modifications to the AIC can be applied to produce an MDL-based estimator.

[Asadi and Seyfe, 2013] takes an altogether different approach based on entropy estimation. Instead of encoding the data with the help of the estimated parameters, the authors estimate the number of signal sources by estimating the entropy of the eigenvalues, and deciding which are driven by the signals and which by the noise.

After the eigenvalues are obtained, their entropy is estimated using kernel smoothing, and the number of signal eigenvalues is determined by the set of eigenvalues with the maximum estimated entropy.

The authors provide simulation results in various scenarios and various types of noise, showing vast improvement over other methods to determine the number of sources. It would be very interesting if using this type on real data will give an improved probability of source detection.

2.7.2. Theory

This section features the recent theoretical advances in the MDL principle.

[Pandey and Dukkipati, 2013] treat the problem of selecting a maximum entropy model given various feature subsets and their moments. The problem is reduced to a model selection problem, for which it turns out that the MDL formulation can be reduced to an NML code. The authors also discuss some connections between the maximum entropy model selection and the minimax entropy principle (when all models are considered of equal complexity).

This paper is of particular interesting because for defined NML models the complexity still has to be calculated, and closed-form solutions exist for few problems. The culprit here is that the maximum-entropy distribution that has the given feature subsets is an exponential family:

$$\int_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x})\phi_k(\mathbf{x})dx = \overline{\phi}_k(\mathbf{x}^n)$$

$$p(\mathbf{x}) = \exp\left(-\lambda_{0,l} - \sum_{k=1}^{m_l} \lambda_{k,l}\phi_{k,l}(\mathbf{x})\right)$$

The resulting complexity is

$$COMP(\mathcal{M}_\Phi) = \log \int_{y^n\in\mathcal{X}^n} \exp\left(-nH(p^*_{y^n})\right) d\mathbf{y}^n$$

where $p^*_{y^n}$ is the maximum entropy distribution for $\mathcal{L}_{\Phi,y^n}$.

The case study is gene identification - from known classification of subjects the genes which determine the outcomes are compressed using feature identification and then

used to compress the classification of the cases. Some details on the method of application is not clear.

The next cluster of articles talks about the connections between NML, a related model - Sequential NML (SNML) and Bayesian universal models with Jeffreys' prior.

[Kotlowski and Grunwald, 2011] discuss the relative merits of the NML and the SNML universal models - naturally a problem for the NML is the fact that it depends a lot on the estimated horizon, so is not very useful when the data from process that is being modeled arrives sequentially.

SNML was developed in [Rissanen and Roos, 2007, Roos and Rissanen, 2008] in order to work with incremental increase of the sample size. The question however remained whether the performance of SNML is significantly different than the NML.

In this regard [Kotlowski and Grunwald, 2011] proves that thee SNML asymptotically behaves just like the NML, in that for exponential families, under certain regularity conditions, the model complexity is asymptotically equal to

$$\frac{k}{2}\log n + O(1). \tag{2.14}$$

This is a strong result, because the NML is optimal (in the sense of minimax regret), so the SNML is also asymptotically optimal.

The article poses some interesting questions, one of which is of the relation between the SNML and the Bayesian model with Jeffreys' prior. It is known that whenever the sequences x^n are limited to compact subsets of $\mathcal{X}^n$, the Bayes with Jeffreys' prior are asymptotically optimal, i.e. achieves the same regret as the NML, ((2.14)).

The question is what happens when we can relax the conditions - are the SNML and the Bayes with Jeffreys' prior are the same, and what is the relation between the minimax regret when they are not? These problems are tracked by the next two papers.

In the first paper [Hedayati and Bartlett, 2012b] prove the major result that for canonical exponential families, the condition on the distribution that is necessary and sufficient to have the SNML, NML and Bayesian strategy with Jeffreys' prior be equivalent, is that the distribution p_{SNML} is exchangeable. That is,

$$p(x^{m..n}|x^{m-1}) = p(x^{\pi(m..n)}|x^{m-1}), \tag{2.15}$$

for all permutations π. This means reordering the sample (as long as we preserve the part on which we are conditional, x^{m-1}) will not change the conditional probability.

A corollary of the above is that when the SNML and NML are equivalent, the strategy for the NML becomes independent of n.

Note that this precludes the direct use of many time-series models, because time dependence is not compatible with (2.15).

Later in [Hedayati and Bartlett, 2012a] the question moves to a more general setup, where the families are not necessarily Gaussian. The regularity condition required in this case for the major results to hold is that the ML estimator is asymptotically Gaussian. The authors also turn special attention to the quantization effects that arise in real-world application of the MDL.

The main result in [Hedayati and Bartlett, 2012a] is that for all parametric models with MLE that is asymptotic normal and has continuous Fisher information exchangeability is equivalent to equivalence between NML, Bayesian with Jeffreys' prior and SNML.

This leaves still leaves open the question if normality is necessary condition for Jeffreys' to asymptotically coincide with SNML.

In summary, from [Kotlowski and Grunwald, 2011, Hedayati and Bartlett, 2012b, Hedayati and Bartlett, 2012a] we have learned that a Bayesian with Jeffreys' prior, NML and SNML are even more intimately linked than before, and strong regularity conditions can be weakened and still they are shown to be asymptotically equivalent.

More works need to be done on how finite-sample performance of the SNML, NML and Bayes with Jeffreys' prior compare to each other. After all, data in many important practical problems can be very limited, so asymptotic considerations may not apply.

In this connection, the construction of optimal plug-in codes (in term of redundancy rates) is tracked in [Grünwald and Kotlowski, 2010]. It solves the posed general open problem in [Grunwald, 2007] if there is a way to make plug-in codes optimal, at least in asymptotic sense.

The answer turns out to be "yes", and the code in question is the squashed ML code, which unlike the ordinary prequential ML code modifies the estimated code by allowing a bit more variability of the future data. The idea is similar to Bayesian estimation of parameters, effectively "unbiasing" the estimates, also similar to unbiasing the variance by dividing by $n-1$ instead of n for Gaussian with unknown mean.

Yet another paper which connects much to the above is [Harremoës, 2013]. The authors work through several characterizing theorems on the connection between the Jeffreys' prior and the NML. They prove that the MDL principle combined with the exchangeability condition leads directly to Bayesian models with Jeffreys' prior, including the case where Jeffreys' prior is improper, with notes on the treatment.

Another interesting point is the philosophical discussion about the nature of the ML, primarily a frequentist concept, and the Bayesian models employed, and how MDL comes between them. The idea is that in order to justify use of idealized codelength via Kraft's inequality, we must consider data sequences as part of a longer sequence (Bayesian approach does not require this), but this longer sequence can be finite, unlike the frequentist approach which requires the sequence be a part of an infinite sequence in order to be theoretically sound.

The treatment is done for exponential families only, like in the already mentioned [Kotlowski and Grunwald, 2011, Hedayati and Bartlett, 2012b], which the authors claim is due to technical reasons - it would be interesting to see the general treatment for different parametric families like the one in [Hedayati and Bartlett, 2012a].

The next cluster of articles [Nonchev, 2013b, Nonchev, 2013a, Nonchev, 2014] focuses on the NML universal model, in particular to its application to model selection in term of its distribution.

A different statistic, called *distribution complexity* (DC), derived from the *model complexity*, is introduced in [Nonchev, 2013b] for scale-location families of distributions. The distribution complexity can be used for model selection to determine the marginal distribution, and in particular deals with the problem of with the infinite model complexity.

The next article [Nonchev, 2013a] deals with the distribution complexity of spherical distributions, generated by a univariate distribution. The major result is that the distribution complexity for all spherical distributions is analytically computed, but the combined description length is the same regardless of the concrete spherical distribution, which means that according to the MDL principle they are indistinguishable and encode samples with the same length of the code.

The last article [Nonchev, 2014] lays the foundations for optimized numerical approach in the computation of the distribution complexity for different distributions, with particular interest in fat-tailed distributions.

2.7.3. Philosophy

A very interesting critique on the theoretical limitations of the MDL principle is expounded in [Adriaans and Vitanyi, 2007]. The authors define a very basic definition of the MDL principle, in language more compatible with Kolmogorov complexity, namely Kolmogorov sufficient statistic (introductory material is available in Section A.6 or in Chapter 14, [Cover and Thomas, 2012]).

The following points are discussed in the paper:

- computability - like Kolmogorov complexity, the structure function (MDL) is not computable; thus
- MDL is computed iteratively, but cannot stop in the proper place, since it is not known when the minimum description is achieved; moreover
- shorter MDL code does not necessarily mean a better model (unless we have found really the best, which there is no way to know).

The example presented constitutes of inferring a grammar from positive examples, for which the authors present an MDL code. For some data samples, successive iterations reduce the description length, but nevertheless do increase the randomness deficiency.

It is necessary to point out that this critique/limitations of the MDL may be interpreted as justification for the necessary trade-off - in some regards it is not entirely fair to measure MDL codes with arbitrary computer codes, the shortest of which are defined as the Kolmogorov complexity.

Still it is necessary to be reminded that the shorthand formulation of the MDL principle that shorter codes lead to better models, is just a shorthand, and the precise definition of the MDL principle is that the best explanation is the one that produces the shortest code. This paper proves that how we arrive at the shortest description is not always clear, and if we can really achieve it is not always provable.

The next article [Zenil and Soler-Toscano, 2012] deals with Kolmogorov complexity for two-dimensional objects.

As mentioned in Chapter 2 the inspiration for the Minimum Description Length principle has historically been the Kolmogorov complexity, introduced by Kolmogorov in his seminal article [Kolmogorov, 1963]. The main drawbacks of using the complexity directly is that it is not computable, and it can be arbitrary for short-length data samples.

In [Zenil and Soler-Toscano, 2012] the authors mitigate somewhat the problem of arbitrariness by restricting the set of considered turmites (two-dimensional Turing machines) to absolute turmites, and encode the exact turmite with an explicit coding scheme. The computability issue is sidestepped, because, as most of the turmites halt quickly, the minimum program length can be calculated directly most of the time. Thus with comparatively short run-time the Kolmogorov complexity can be approximated quite well.

The interesting problem that the authors try to solve is to demonstrate that the approximation of Kolmogorov complexity by algorithmic complexity is consistent with the approximation done by compression.

Another major point is that the problem posed in the paper is for multi-dimensional Kolmogorov complexity, which has distinct features and is not clearly approximated by Shannon codes tuned for one-dimensional patterns, see [Andrienko et al., 2000].

Another notable point is how the authors track the problem of compression of small samples - by concatenating the small objects and encoding them en masse to obtain fractional approximation of the coding length.

2.7.4. Conclusion

More than 35 years after [Rissanen, 1978], Rissanen's seminal paper on the MDL principle, the field is still growing to be more and more relevant to real-world applications.

The theoretical elegance of the MDL principle is combined with the capacity to provide clear and unambiguous criteria for unsupervised learning without the need

for fine-tuning parameters, and even more strongly, the possibility to provide a method for estimation of such parameters.

A point that needs further improvement is the application of universal models, as many practical applications still use two-part codes, even default to the "crude" MDL principle, requiring explicit coding of the hypothesis. The choice of such codes is not always practical or the preference of one code over another justifiable.

On the theoretical side there is more need to explore the limitations of the universal models. More work is required on the model complexity, in particular the infinity problem - practically every non-trivial NML code leads to infinite model complexity, which either prevents its use or requires imposing artificial limitations on the model.

There is also more need for small-sample comparisons and connections, for example between the Bayesian universal model and the NML. Even in the age of big data a lot of statistical inference problems have limited sample size, and the MDL principle is well-suited for them.

3. Distribution Complexity

This chapter describes the main contribution of this dissertation. It is dedicated to the creation of a usable MDL-based method to distinguish and choose between different distributions, particularly those that have fat tails, against a basic, Gaussian distribution.

The goal of *distribution complexity* is, by serving as a solution to the infinity problem from Section 2.6.4, to provide a way to apply the Stochastic Complexity criterion to the problem of selecting the distribution of a sample.

This chapter is based on [Nonchev, 2013b, Nonchev, 2013a]. The notion of distribution complexity was introduced in [Nonchev, 2013b], and the analytic formula for the complexity of spherical distributions first presented in [Nonchev, 2013a]. The general case of independent samples can be found in [Nonchev, 2014].

As a motivation problem we focus on two multivariate distributions, which are assumed uncorrelated. This implies independence only for the Gaussian model.

For uncorrelated distributions, the distribution complexity has a closed-form expression that we discuss in Section 3.3.

For independent samples, i.e. whose univariate distributions are independent, we extend the formula of distribution complexity and discuss the theoretical results needed to create algorithm for efficient calculation of the distribution complexity in Section 3.4.

The next step in the extension of the results to inclusion of additional parameters, called shape parameters, is done in Section 3.5.

This chapter is concerned only with the theoretical advances. The numerical computations based on the formulas herein are discussed in details in Chapter 4.

3.1. Motivation

This section introduces the models of interest that serve as motivating example for the distribution complexity.

When talking about the distribution of a sample in statistics, we usually assume that the sample is derived from i.i.d. experiments. The natural candidates in such a setting are exponential families, which is a strong limitation. This is a very

convenient assumption, but thwarts the variety of distributions that are analytically tractable.

For our purposes we want to analyze the effects that the shape of the distributions, we introduce the distributions for the whole sample, which is not standard practice in statistics.

In terms of the Refined MDL principle from Section 2.5, we want to find the probabilistic source from a set of possible distributions that best explains the given sample.

3.1.1. Models of Interest

As a motivation problem we focus on two multivariate distributions, which are assumed uncorrelated. As already mentioned, this implies independence only for the Gaussian model.

First, some definitions.

Definition 3.1 (Uncorrelated Gaussian model)**.** The first model considered for the sample is a multivariate Gaussian distribution with p.d.f.:

$$f^N_{\mu,\sigma}\left(x^n\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)\right).$$

The elements of such a sample are also independent, and this is a characterizing property of the multivariate Gaussian distribution.

Definition 3.2 (Uncorrelated Student-T distribution)**.** The second model considered for the sample is a multivariate uncorrelated Student-T distribution with fixed degrees of freedom ν_0, having p.d.f.

$$f^T_{\mu,\sigma}\left(x^n\right) = \frac{\Gamma\left(\frac{n+\nu_0}{2}\right)}{\Gamma\left(\frac{\nu_0}{2}\right)\left(\nu_0\pi\sigma^2\right)^{\frac{n}{2}}}\left(1+\frac{1}{\nu_0}\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right)^{-\frac{n+\nu_0}{2}}.$$

The reasons behind the choice of Student-T distribution is that it is quite similar to the Gaussian distribution, although when ν_0 is small, it exhibits distinctly heavy tails.[1] Asymptotically for $\nu_0 \to \infty$, the Student-T converges to Gaussian.

Note however that unlike in a frequentist hypothesis testing, we do not consider either of the models above to be the default choice of model, and the other as

[1]The term *heavy tail* usually refers to distributions that have substantially heavier tails than the Gaussian. The classical example of such a distribution is the Levy α-stable distribution, which has much heavier tails than the Student-T in general.

alternative - that necessitates more knowledge that we care to admit. Nor do we assign prior probabilities of the models as in a Bayesian setting (see Section 1.1.1.3).

As mentioned above, the individual observations in the Gaussian model are independent, in addition to identically distributed. In the Student-T model they are only uncorrelated, so that model cannot be used to model an IID sample (see [Kotz and Nadarajah, 2004], Chapter 1).

Otherwise, this definition of the multivariate Student-T distribution has the advantage that the maximum likelihood estimator for the parameters μ and σ are, as in the case of the Gaussian distribution, the sample mean and variance. This characterizes the so-called spherical distributions, which will be explored further in

In addition the MLE of a linear regression is the ordinary least squares method (see [Kotz and Nadarajah, 2004], Chapter 11). This means that the discussed Student-T model can easily replace the Gaussian distribution as the noise distribution in a linear regression.

3.1.2. Stochastic Complexity of the Gaussian Distribution

There is a classical result for the complexity of the Gaussian model, and it is presented in this section following [Rissanen, 2000]. The known ways to deal with infinite model complexity are also addressed below.

To do so, we recall the notion of sufficiency, presented in Section 1.1.2. To use it for the Gaussian, recall the following classical result:

Lemma 3.3. The sample estimates (see Definition 1.14) of the mean $\overline{x}$ and variance s_x^2 are sufficient statistics for the Gaussian model.

Moreover, for the Fisher-Neyman factorization ((1.3)) the dependence between the statistics and the parameters h has the form

$$g_{\mu,\sigma}\left(\overline{x},s\right)=\frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)\frac{\sigma^2}{n}}\left(\frac{ns^2}{\sigma^2}\right)^{\frac{n-1}{2}-1}e^{-\frac{ns^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi\left(\frac{\sigma}{\sqrt{n}}\right)^2}}\exp\left[-\frac{\left(\overline{x}-\mu\right)^2}{2\left(\frac{\sigma}{\sqrt{n}}\right)^2}\right]. \quad (3.1)$$

Proof. Let x^n be an i.i.d. sample of Gaussian distributed random variables from $N\left(\mu,\sigma^2\right)$. The joint distribution of the sample is expressible as

$$\begin{aligned} f_{\mu,\sigma}^N\left(x^n\right) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left(-\frac{1}{2\sigma^2}\sum\left(x_i-\mu\right)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left(-\frac{ns^2}{2\sigma^2}\right)\exp\left(-\frac{\left(\overline{x}-\mu\right)^2}{2\left(\frac{\sigma}{\sqrt{n}}\right)^2}\right). \end{aligned} \quad (3.2)$$

Since the dependence in (3.2) is only to $\overline{x}$ and s^2, this shows that they are sufficient statistics, so we have to only find the Fisher-Neyman factorization (Theorem 1.19). We will use the following form:

$$f_{\mu,\sigma}\left(x^n\right) = h\left(x^n\right) g_{\mu,\sigma}\left(\overline{x}, s\right),$$

where $g_{\mu,\sigma}\left(\overline{x}, s\right)$ is the p.d.f. of distribution of x and s^2.

Cochran's theorem provides that when X_i are i.i.d. and Gaussian, $\overline{x} \in \mathbb{N}\left(\mu, \frac{\sigma^2}{n}\right)$ and $\frac{ns^2}{\sigma^2} \in \chi^2_{n-1}$, and both are independent of each other. Thus we arrive at the joint p.d.f. of the sufficient statistics in (3.1). □

We cite the following result for model complexity when the maximum likelihood estimator is itself a sufficient statistic. It is available for example in [Barron et al., 1998], Chapter III, Section F. We will generalize it for constrained complexity.

Theorem 3.4 (Constrained complexity for sufficient statistic)**.** *If the MLE $\hat{\theta}$ for a parametric family $\mathcal{M}_\theta$ is a sufficient statistic, then the constrained model complexity, conditioned on $x^n \in \left\{x^n : \hat{\theta}(x^n) \in \mathcal{A}\right\}$ is expressible as*

$$\exp COMP\left(\mathcal{M}|\hat{\theta}(x^n) \in \mathcal{A}\right) = \int_{\hat{\theta}(x^n)\in\mathcal{A}} f_{\hat{\theta}(x^n)}\left(x^n\right) dx^n = \int_{t\in\mathcal{A}} g_t\left(t\right) dt,$$

where $g_\theta\left(t\right)$ is the probability density of the sufficient statistic for the given value of θ.

Proof. The MLE being a sufficient statistic is equivalent to the existence of the following Fisher-Neyman decomposition:

$$f_\theta\left(x^n\right) = h(x^n) g_\theta(\hat{\theta}(x^n)),$$

where according to Theorem 1.20 we can choose g_θ to be the p.d.f. of $\hat{\theta}$. In this case $h(x^n)$ is the conditional probability distribution, i.e. of $h(x^n)$, over the condition that $\hat{\theta}$ has specific value.

To complete the proof we can use the δ-calculus (see Appendix B) and expand the

integral as

$$\begin{aligned}
\exp COMP\left(\mathcal{M}|\hat{\theta}(x^n) \in \mathcal{A}\right) = & \int_{\hat{\theta}(x^n)\in\mathcal{A}} f_{\hat{\theta}(x^n)}\left(x^n\right) dx^n \\
= & \int_{t\in\mathcal{A}} \int_{x^n\in\chi^n} f_t\left(x^n\right) \delta(\hat{\theta}(x^n) - t) dx^n dt \\
= & \int_{t\in\mathcal{A}} \int_{x^n\in\chi^n} h(x^n) g_t(\hat{\theta}(x^n)) \delta(\hat{\theta}(x^n) - t) dx^n dt \qquad (3.3) \\
= & \int_{t\in\mathcal{A}} g_t(t) \left[\int_{x^n\in\chi^n} h(x^n) \delta(\hat{\theta}(x^n) - t) dx^n\right] dt \qquad (3.4) \\
= & \int_{t\in\mathcal{A}} g_t(t) dt. \qquad (3.5)
\end{aligned}$$

In the above equation (3.3) follows from the assumption of sufficiency, and (3.5) follows from Theorem 1.20, because $h(x^n)$ is a conditional probability distribution (see Corrolary 1.21). □

A particular case when $\mathcal{A} = \Theta$ will provide the following

Corollary 3.5 (Complexity for sufficient statistic)**.** *If the MLE $\hat{\theta}$ for a parametric family $\mathcal{M}_\theta$ is a sufficient statistic, then the model complexity is expressible as*

$$\exp COMP\left(\mathcal{M}\right) = \int f_{\hat{\theta}(x^n)}\left(x^n\right) dx^n = \int_{s\in\Theta} g_s\left(s\right) ds. \qquad (3.6)$$

As described in detail in Section 2.6.4 the above integral is infinite in our model, i.e. the complexity is not well defined, so the case covered by Corrolary 3.5 is not sufficient.

Also note that this integral is also hard to solve for large n using direct numerical integration, because it is n-dimensional, so a better approach is needed. It is covered in Section 4.2 below.

There are several approaches to deal with this that were outlined in Section 2.6.4. Most notable are the "renormalization" by complexity conditional on the data space (constrained complexity) as presented in [Rissanen, 2000] and the usage of complexity conditional on the parameter space as in [Stine and Foster, 2001].

Both approaches have their merits. However, to compare between models we have to account for the various parameterizations, thus limiting on the sample space x^n using Definition 2.27 allows for more flexibility.

We finish the motivation section with the classical result for constrained complexity for the Gaussian.

Theorem 3.6. *Let $M_{\mu,\sigma}$ be a Gaussian family of distributions. The constrained complexity (Definition 2.27) on*

$$\mathcal{A} = \{x^n : \hat{\mu}(x^n) \in [-R; R] \cap \hat{\sigma}(x^n) \in [D, \infty)\}$$

is then expressed as

$$\exp COMP(\mathcal{M}|x^n \in \mathcal{A}) = 2RD^{-1}\frac{2n^{\frac{n}{2}}e^{-\frac{n}{2}}}{\sqrt{2\pi}2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}.$$

Proof. Substituting $\mu = \overline{x}$ and $\sigma^2 = s^2$ into (3.1) we get

$$g_{\mu=\overline{x},\sigma=s}\left(\overline{x}, s^2\right) = \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)\frac{s^2}{n}}(n)^{\frac{n-3}{2}}e^{-\frac{n}{2}}\frac{\sqrt{n}}{\sqrt{2\pi s^2}}.$$

Then we apply this to Theorem 3.4 to obtain an analytic formula for the constrained complexity :

$$\begin{aligned}
\exp COMP\left(\mathcal{M}|\hat{\theta}(x^n) \in \mathcal{A}\right) &= \int_D^\infty \int_{-R}^R \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)\frac{s^2}{n}}(n)^{\frac{n-3}{2}}e^{-\frac{n}{2}}\frac{\sqrt{n}}{\sqrt{2\pi s^2}}dm\,ds^2 \\
&= 2R\frac{n}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}(n)^{\frac{n-3}{2}}e^{-\frac{n}{2}}\frac{\sqrt{n}}{\sqrt{2\pi}}\int_D^\infty (s^2)^{-3/2}\,ds^2 \\
&= 2RD^{-1}\frac{2n^{\frac{n}{2}}e^{-\frac{n}{2}}}{\sqrt{2\pi}2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}. \qquad (3.7)
\end{aligned}$$

□

3.1.3. Dealing with Infinite Complexity

The idea utilized in this dissertation to tackle the infinite complexity is to separate the terms and treat them differently. If we cancel the common terms between two distributions, we can compare their complexity for fixed R and D, so the comparison does not actually dependent on the boundaries, even thought the model complexity depends on them.

Moreover, the same relationship will hold for all R and D. To do that we need a particular decomposition of the complexity of a scale-location family.

The term of interest is $2RD^{-1}$. Since it is a multiplicative constant, we can isolate it as a separate in the complexity:

$$COMP\left(\mathcal{M}|\hat{\theta}(x^n) \in \mathcal{A}\right) = \ln 2 + \frac{n}{2}\left(\ln\frac{n}{2} - 1\right) - \frac{\ln \pi}{2} - \ln\Gamma\left(\frac{n-1}{2}\right) + \ln 2RD^{-1}.$$

In papers like [Rissanen, 2000] and [Stine and Foster, 2001], the codes are extended to encode the limits R and D, but this introduces arbitrariness as some parameter values are treated as more likely than others.

The approach in this dissertation is to use complexity conditional on the data space (constrained complexity), but without re-normalization, as it is not needed when comparing the two chosen models, and consequently no arbitrariness arises.

Fact 3.7. *The last term* $\ln 2RD^{-1}$ *does not depend on the sample size, and captures all of the dependence on the boundaries of integration. The rest of the terms capture the model complexity.*

Thus we justify the application of Stochastic Complexity (SC) criterion (Section 2.6.3) across model classes such as different location and scale families by subtracting the common term $\ln 2RD^{-1}$.

To take advantage of that fact, we turn over to general scale-location families in Section 3.2.

3.2. Scale-Location Families

As mentioned in Section 3.1 we work with the entire sample x^n, so the definition of scale-location families is given in the multivariate case. This more general definition permits non-independently distributed samples.

Definition 3.8. A scale-location family is a family of distributions having p.d.f. $f_{\mu,\sigma}(x^n)$ for which a distribution $f(x^n)$ exists satisfying

$$f_{\mu,\sigma}(x^n) = \sigma^{-n} f\left(\frac{x^n - \mu}{\sigma}\right)$$

The standard member of the distribution $f(x^n)$, i.e. the one having $\mu = 0$ and $\sigma = 1$, will feature more prominently in the analysis.

The following lemma shows an important characterization of the scale-location families and their corresponding maximum likelihood estimators, a well-known fact. The proof is also provided for completeness.

Lemma 3.9. If $\hat{\mu}(x^n)$ and $\hat{\sigma}(x^n)$ are MLE for a scale-location family (i.e. they exist and are unique), then for any $a \in \mathbb{R}^+$ and $b \in \mathbb{R}$ we have

$$\hat{\mu}(ax^n + b) = a\hat{\mu}(x^n) + b$$

$$\hat{\sigma}(ax^n + b) = a\hat{\mu}(x^n).$$

Proof. By the definition of a MLE (Definition 1.16) we have

$$\left(\hat{\sigma}\left(x^n\right), \hat{\mu}\left(x^n\right)\right) = \arg_{\mu,\sigma}\max f_{\mu,\sigma}\left(x^n\right) = \arg_{\mu,\sigma}\max \sigma^{-n} f\left(\frac{x^n-\mu}{\sigma}\right).$$

Using the definition of the p.d.f. of a scale-location family, we get

$$\begin{aligned}\left(\hat{\sigma}\left(ax^n+b\right), \hat{\mu}\left(ax^n+b\right)\right) &= \arg_{\mu,\sigma}\max f_{\mu,\sigma}\left(ax^n+b\right)\\ &= \arg_{\mu,\sigma}\max \sigma^{-n} f\left(\frac{ax^n+b-\mu}{\sigma}\right)\\ &= \arg_{\mu,\sigma}\max \left({}^{\sigma}/_{a}\right)^{-n} f\left(\frac{x^n+(b-\mu)/a}{\sigma/a}\right).\end{aligned}$$

Combined with the definition, we get the following equities

$$\begin{aligned}\hat{\sigma}\left(ax^n+b\right)/a &= \hat{\sigma}\left(x^n\right)\\ \left(\hat{\mu}\left(ax^n+b\right)-\mu\right)/a &= \hat{\mu}\left(x^n\right).\end{aligned}$$

□

Now we introduce a more succinct way to talk about the model complexity for scale-location families without running into infinite model complexity.

Definition 3.10 (Distribution complexity)**.** The distribution complexity is defined as

$$DC_n\left(\mathcal{M}\right) = \log \mathbb{E}_{X^n}\left[\delta\left(\hat{\mu}\left(X^n\right)\left(1-\hat{\sigma}\left(X^n\right)\right)\right)\right]. \tag{3.8}$$

We will use the Dirac δ-function for brevity of notation, and prove the following major result.

Theorem 3.11 (Model complexity from distribution complexity)**.** *For a scale-location family the constrained complexity based on*

$$\mathcal{B} = \{x^n : |\hat{\mu}(x^n)| \leq R, \hat{\sigma}(x^n) \geq D\}$$

can be decomposed as

$$COMP\left(\mathcal{M}|x^n \in \mathcal{B}\right) = \log 2RD^{-1} + DC_n\left(\mathcal{M}\right). \tag{3.9}$$

Proof. The first step is to rewrite the integral using the standard density $g(x^n)$:

$$
\begin{aligned}
& \exp COMP \\
= & \int_{x^n \in \mathcal{B}} f_{\hat{\mu}(x^n),\hat{\sigma}(x^n)}\left(x^n\right) dx^n \\
= & \int_{x^n \in \mathcal{B}} \int\int \delta\left(\mu - \hat{\mu}\left(x^n\right)\right) \delta\left(\sigma - \hat{\sigma}\left(x^n\right)\right) f_{\mu,\sigma}\left(x^n\right) d\mu \, d\sigma \, dx^n \\
= & \int_{x^n \in \mathcal{B}} \int\int \delta\left(\mu - \hat{\mu}\left(x^n\right)\right) \delta\left(\sigma - \hat{\sigma}\left(x^n\right)\right) \sigma^{-n} f\left(\frac{x^n - \mu}{\sigma}\right) d\mu \, d\sigma \, dx^n
\end{aligned}
$$

Now let us turn our attention to $\mathcal{B}$, so that we can move the boundaries of the integral from $\hat{\mu}$ and $\hat{\sigma}$ to μ and σ. Since $\mathcal{B}$ is defined as those x^n for which $\hat{\mu}\left(x^n\right) \in [-R; R]$ and $\hat{\sigma}\left(x^n\right) \in [D, \infty)$, and the δ-function has support only $\{0\}$, we can change the inner integral limits to $[-R; R]$ and $[D; \infty)$:

$$
= \int_{x^n \in \mathbb{R}^n} \int_{-R}^{R} \int_{D}^{\infty} \delta\left(\mu - \hat{\mu}\left(x^n\right)\right) \delta\left(\sigma - \hat{\sigma}\left(x^n\right)\right) \sigma^{-n} g\left(\frac{x^n - \mu}{\sigma}\right) d\mu d\sigma dx^n \tag{3.10}
$$

We make the substitution $y^n = \frac{x^n - \mu}{\sigma}$, for which $|J| = \sigma^n$. Since $\hat{\mu}\left(x^n\right)$ and $\hat{\sigma}\left(x^n\right)$ are MLE for a scale-location family we can use the Lemma lemma 3.9 to simplify (3.10). Combining that with the fact that δ is homogeneous of degree -1 (see lemma B.8), we can isolate the dependence on the boundaries and σ and μ as follows:

$$
\begin{aligned}
= & \int_{y^n \in \mathbb{R}^n} \int_{-R}^{R} \int_{D}^{\infty} \delta\left(\mu - \sigma\hat{\mu}\left(y^n\right) - \mu\right) \delta\left(\sigma - \sigma\hat{\sigma}\left(y^n\right)\right) \sigma^{-n} f\left(y^n\right) \sigma^n d\mu \, d\sigma \, dy^n \\
= & \int_{y^n \in \mathbb{R}^n} \int_{-R}^{R} \int_{D}^{\infty} \sigma^{-2} \delta\left(\hat{\mu}\left(y^n\right)\right) \delta\left(1 - \hat{\sigma}\left(y^n\right)\right) f\left(y^n\right) d\mu \, d\sigma \, dy^n \\
= & \left[\int_{-R}^{R} \int_{D}^{\infty} \sigma^{-2} d\mu \, d\sigma\right] \int_{y^n \in \mathbb{R}^n} \delta\left(\hat{\mu}\left(y^n\right)\right) \delta\left(1 - \hat{\sigma}\left(y^n\right)\right) f\left(y^n\right) \, dy^n \\
= & \, 2RD^{-1} \int_{y^n \in \mathbb{R}^n} \delta\left(\hat{\mu}\left(y^n\right)\right) \delta\left(1 - \hat{\sigma}\left(y^n\right)\right) f\left(y^n\right) \, dy^n.
\end{aligned}
$$

So we have arrived at the quantity of interest

$$
\begin{aligned}
\frac{\exp COMP\left(\mathcal{M}|\mathcal{B}\right)}{2RD^{-1}} &= \int_{y^n \in \mathbb{R}^n} \delta\left(\hat{\mu}\left(y^n\right)\right) \delta\left(1 - \hat{\sigma}\left(y^n\right)\right) f\left(y^n\right) \, dy^n \\
&= \int_{y^n \in \mathbb{R}^n} \delta\left(\hat{\mu}\left(y^n\right)\right) \delta\left(1 - \hat{\sigma}\left(y^n\right)\right) \, dF\left(y^n\right) \\
&= \mathbb{E}_{X^n}\left[\delta\left(\hat{\mu}\left(X^n\right)\left(1 - \hat{\sigma}\left(X^n\right)\right)\right)\right] \qquad (3.11) \\
&= \exp DC_n\left(\mathcal{M}\right).
\end{aligned}
$$

□

The derivation above is the reason we call $DC_n(\mathcal{M})$ the distribution complexity, as it depends only on the shape of the marginal distribution and the dependence structure, and not on the restriction of the parameters.

This also introduces us to the basic fact that for all scale-location families we are going to experience the same problem with infinite complexity.

Corollary 3.12. *For scale-location families the non-constrained model complexity is always infinite.*

Proof. Use Theorem 3.11 and let $R \to \infty$ and $D \to 0$ in (3.9). □

The SC criterion (Section 2.6.3) in the special case of selecting between our two models is rewritten as

$$\tilde{L}_{DC}(x) = \ln f_{NML}(x) - \ln 2RD^{-1} = -\ln f_{\hat{\mu}(x),\hat{\sigma}(x)}(x) + DC_n(\mathcal{M})$$

and comparison is done using the adjusted codelength $\tilde{L}_{DC}(x)$.

3.2.1. Stochastic Complexity Criterion Revisited

Before continuing to calculate the distribution complexity, we will reformulate the Stochastic Complexity criterion from Section 2.6.3 the following way:

Definition 3.13 (Stochastic Complexity criterion with Distribution Complexity)**.** Let $\mathcal{M}_1, \mathcal{M}_2, \ldots$ be a set of possible models corresponding to scale-location families for a sample x^n. The description length of x^n for each model can be defined as

$$SC(\mathcal{M}_i) = -\log P_{\mathcal{M}_i}(y^n|\hat{\theta}(y^n)) + DC(\mathcal{M}_i).$$

The best model is the one with minimum $SC(\mathcal{M}_i)$.

3.3. Spherical scale-location families

Uncorrelated samples introduce the so-called spherical families. They are families for which the probability depends only on the distance from some location μ, compared to a scale parameter σ. Since the samples are identically distributed, the location and scale are the same for all elements.

First we start by the definition of spherical distributions.

Definition 3.14 (Spherical distribution)**.** Suppose we have an arbitrary univariate distribution satisfying $f(x) = cK(x)$ for an even function K.

The multivariate spherical generalization of f is defined as

$$f_{\mu,\sigma}(x^n) = c\sigma^{-n} K\left(\sigma^{-2}\left(x^n-\mu\right)^T\left(x^n-\mu\right)\right), \tag{3.12}$$

where c is the normalizing constant of proportionality.

K is further down referred to as the generator of the spherical distribution.

Note that c is constant with respect to μ and σ. To see that just do a transformation $(x^n) \to (y^n = \frac{x^n-\mu}{\sigma^n})$ and isolate the effect of K by the following

$$\begin{aligned}(c\sigma^{-n})^{-1} &= \int K\left(\sigma^{-2}\left(x^n-\mu\right)^T\left(x^n-\mu\right)\right) dx^n \\ &= \int K\left(y^n\right)|J| dy^n \\ &= \sigma^n \int K\left(y^n\right) dy^n,\end{aligned}$$

therefore

$$c = \left(\int K\left(y^n\right) dy^n\right)^{-1}.$$

The defining feature of a spherical family is that $\overline{x} = \frac{1}{n}\sum x_i$ and $s^2 = \frac{1}{n}\sum x_i^2$ are sufficient statistics for μ and σ. To use them we introduce mild regularity conditions and prove the following elementary lemma.

Definition 3.15. The following conditions will be referred as *regularity conditions* for the generator K:

1. $\frac{\partial K}{\partial w}$ exists for $\forall w > 0$;
2. $\frac{\partial K}{\partial w} < 0$ for $\forall w > 0$;
3. one of the solutions of $\frac{\partial \log K}{\partial w}(w) = -\frac{n}{2w}$ attains a maximum of $w^{\frac{n}{2}} K\left(w\right)$.

Lemma 3.16 (Spherical distribution reparameterization)**.** *Let $f_{\mu,\sigma}(x^n)$ be a spherical distribution family and its corresponding generator K. If K satisfies the regularity conditions (Definition 3.15), then there is a reparameterization in which the MLEs of μ and σ correspond to the sample mean and sample standard deviation.*

Proof. We construct the appropriate parmeterization explicitly. Define W_0 to be the set of non-negative solutions of the equation

$$\frac{\partial \log K}{\partial w}(w) = -\frac{n}{2w}.$$

Then we have for $\forall w \in W_0$ that

$$-\frac{2}{n} w \frac{\partial K}{\partial w}(w) = K(w). \tag{3.13}$$

Find $w_0 \in W_0$ that maximizes $w^{\frac{n}{2}} K(w)$, i.e.

$$w_0 = \arg_{w \in W_0} \max w^{\frac{n}{2}} K(w). \tag{3.14}$$

This w_0 sill serve to reparameterize K. Define $\tilde{\sigma}^2 = \frac{n}{w_0}\sigma^2$, therefore

$$K\left(\sigma^{-2}(x^n-\mu)^T(x^n-\mu)\right) = K\left(\frac{w_0}{n}\tilde{\sigma}^2(x^n-\mu)^T(x^n-\mu)\right)$$

and hence

$$\tilde{K}(w) = K\left(\frac{w_0}{n}w\right).$$

$\tilde{K}$ obviously represents the same distribution, so we will just use it instead and denote it by K, and its parameter $\tilde{\sigma}$ by σ.

So, without loss of generality, for the thus defined K, the solution of (3.13) which maximizes (3.14) is $w = n$.

Using (3.12), we will find the necessary and sufficient condition for particular values of μ and σ to be the MLE.

First, taking into account that $w(x^n, \mu, \sigma) = \sigma^{-2}(x^n-\mu)^T(x^n-\mu)$ and find the derivatives of the log-likelihood:

$$\begin{aligned}
\frac{\partial \log f_{\mu,\sigma}(x^n)}{\partial \mu} &= \frac{\partial \log K\left(\sigma^{-2}(x^n-\mu)^T(x^n-\mu)\right)}{\partial \mu} \\
&= \frac{1}{K}\frac{\partial K(\sigma^{-2}(x^n-\mu)^T(x^n-\mu))}{\partial w}\frac{\partial w}{\partial \mu} \\
&= \frac{1}{K}\frac{\partial K(\sigma^{-2}(x^n-\mu)^T(x^n-\mu))}{\partial w}\frac{2n\mu - 2\sum_i x_i}{\sigma^2},
\end{aligned}$$

which is zero if and only if $\mu = \overline{x}$.

$$\begin{aligned}
\frac{\partial \log f_{\mu,\sigma}(x^n)}{\partial \sigma^2} &= -\frac{n}{2}\frac{\partial \log \sigma^2}{\partial \sigma^2} + \frac{1}{K(w)}\frac{\partial K(\sigma^{-2}(x^n-\mu)^T(x^n-\mu))}{\partial w}\frac{\partial w}{\partial \sigma^2} \\
&= -\frac{n}{2\sigma^2} + \frac{1}{K(w)}\frac{\partial K(w)}{\partial w}\frac{-(x^n-\mu)^T(x^n-\mu)}{\sigma^2}\frac{1}{\sigma^2} \\
&= -\frac{1}{\sigma^2}\left[\frac{n}{2} + \frac{1}{K(w)}\frac{\partial K(w)}{\partial w}w\right]
\end{aligned}$$

Let all solutions of

$$\frac{\partial \log f_{\mu,\sigma}(x^n)}{\partial \sigma^2} = 0$$

be denoted as W_0. By construction, $n \in W_0$. Moreover

$$f_{\mu,\sigma}(x^n) = c\sigma^{-n} K\left(\sigma^{-2}\left(x^n - \mu\right)^T\left(x^n - \mu\right)\right) = cs^{-n} w^{\frac{n}{2}} K\left(w\right),$$

and by construction $w = n$ maximizes (3.14), so the M.L.E. is achieved when

$$\sigma^{-2}\left(x^n - \mu\right)^T\left(x^n - \mu\right) = n$$

or, equivalently,

$$\hat{\sigma}^2 = s^2.$$

□

The following elementary corollary is provided for completeness.

Corollary 3.17 (Sufficiency for spherical distributions)**.** *For spherical scale-location families, the sample mean and sample standard deviation are sufficient statistics.*

Proof. Trivial, using the fact that a simple scale reparameterization achieves $\hat{\sigma}^2 = s^2$, $\hat{\mu} = \overline{x}$. □

Before proceeding to prove the main result in this chapter we supply the following lemma, with a slightly roundabout proof of the last piece needed to evaluate the complexity integrals.

Lemma 3.18. *If a Gaussian spherical distribution family $\mathcal{M}_{\mu,\sigma}$ with generator K satisfies*

$$\exp DC_n\left(\mathcal{M}\right) = cK(n) \int_{y^n \in \mathbb{R}^n} \delta\left(\hat{\mu}\left(y^n\right)\right) \delta\left(1 - \hat{\sigma}\left(y^n\right)\right)\, dy^n \tag{3.15}$$

then

$$\int_{y^n \in \mathbb{R}^n} \delta\left(\hat{\mu}\left(y^n\right)\right) \delta\left(1 - \hat{\sigma}\left(y^n\right)\right)\, dy^n = \frac{2n^{\frac{n}{2}} \pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)}.$$

Proof. Recall that from Theorem 3.6 we have that the Gaussian distribution complexity is

$$\exp COMP\left(\mathcal{M}|x^n \in \mathcal{A}\right) = 2RD^{-1} \frac{2n^{\frac{n}{2}} e^{-\frac{n}{2}}}{\sqrt{2\pi} 2^{\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right)}. \tag{3.16}$$

The generator for the Gaussian satisfies

$$c\sigma^{-n}K(\sigma^{-2}\left(x^n-\mu\right)^T\left(x^n-\mu\right)) = (2\pi)^{-\frac{n}{2}}\sigma^{-n}\exp\left\{-\frac{\left(x^n-\mu\right)^T\left(x^n-\mu\right)}{2\sigma^2}\right\}$$

so the constant in (3.15)

$$cK(n) = (2\pi)^{-\frac{n}{2}}\exp\left\{-\frac{n}{2}\right\}.$$

Then putting the together (3.15) and (3.16) we get

$$\frac{2n^{\frac{n}{2}}e^{-\frac{n}{2}}}{\sqrt{2\pi}2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} = (2\pi)^{-\frac{n}{2}}\exp\left\{-\frac{n}{2}\right\}\int_{y^n\in\mathbb{R}^n}\delta\left(\hat{\mu}\left(y^n\right)\right)\delta\left(1-\hat{\sigma}\left(y^n\right)\right)\,dy^n$$

Rearranging the terms, we obtain

$$\begin{aligned}\int_{y^n\in\mathbb{R}^n}\delta\left(\hat{\mu}\left(y^n\right)\right)\delta\left(1-\hat{\sigma}\left(y^n\right)\right)\,dy^n = &\quad \frac{2n^{\frac{n}{2}}e^{-\frac{n}{2}}}{\sqrt{2\pi}2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)(2\pi)^{-\frac{n}{2}}\exp\left\{-\frac{n}{2}\right\}}\\ = &\quad \frac{2n^{\frac{n}{2}}\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)}.\end{aligned}$$

□

With the above reparameterization in mind, the following theorem completely characterizes the distribution complexity of a spherical distribution.

Theorem 3.19 (Distribution complexity of spherical distributions)**.** *Let $\mathcal{M}_{\mu,\sigma}$ be a family of spherical distributions with generator K. Without loss of generality assume that $\sigma^2 = s^2$ and $\mu = \overline{x}$. Then*

$$\exp DC_n\left(\mathcal{M}\right) = \frac{2n^{\frac{n}{2}}\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)}[c\cdot K(n)].$$

Proof. From Definition 3.10 and (3.11) we have that

$$\exp DC_n\left(\mathcal{M}\right) = \int_{y^n\in\mathbb{R}^n}\delta\left(\hat{\mu}\left(y^n\right)\right)\delta\left(1-\hat{\sigma}\left(y^n\right)\right)f\left(y^n\right)\,dy^n. \tag{3.17}$$

From Definition 3.14 we know that for the standard member of $\mathcal{M}_{\mu,\sigma}$ we have

$$f(x^n) = cK\left(\left(x^n\right)^T x^n\right),$$

and because the integration in (3.17) happens only for y^n for which $s^2 = \hat{\sigma} = 1$ and $\overline{x} = \hat{\mu} = 0$ we have that

$$f(y^n) = cK(n)$$

Then we can rewrite (3.17) as

$$\exp DC_n(\mathcal{M}) = cK(n) \int_{y^n \in \mathbb{R}^n} \delta(\hat{\mu}(y^n)) \, \delta(1 - \hat{\sigma}(y^n)) \, dy^n.$$

Now, applying lemma 3.18 we get that

$$\exp DC_n(\mathcal{M}) = cK(n) \frac{2n^{\frac{n}{2}} \pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)}$$

□

Note the following corollary from lemma 3.18:

Corollary 3.20. *The following integral evaluates as*

$$\int_{y^n \in \mathbb{R}^n} \delta(\hat{\mu}(y^n)) \, \delta(1 - \hat{\sigma}(y^n)) \, dy^n = \frac{2n^{\frac{n}{2}} \pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)}.$$

Proof. Using lemma 3.18 and noting that

- the integral does not depend on the generator;
- the formula (3.17) is proven for Gaussian distribution.

□

Therefore, for any spherical distribution satisfying the above weak conditions we have analytic formula for the distribution complexity.

We will consider the application of Theorem 3.19 to model selection, in particular the Stochastic Complexity criterion (see Definition 2.26).

Corollary 3.21. *Let $\mathcal{M}_{\mu,\sigma}$ be a spherical distribution scale-location family. The description length in the Stochastic Complexity criterion (Definition 2.26) does not depend on the generator K, and is expressed as*

$$L_{NML}^{\mathcal{M}_{\mu,\sigma}}(x^n) = n \ln s_x + \ln \frac{2n^{\frac{n}{2}} \pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)}$$

Proof. Substitute the likelihood evaluated at the MLE:

$$f_{\mu=\hat{\mu},\sigma^2=\hat{\sigma}^2}(x^n) = f_{\mu=\overline{x},\sigma^2=s_x^2}(x^n) = cs_x^{-n}h\left(s_x^{-2}\left(x^n-\overline{x}\right)^T\left(x^n-\overline{x}\right)\right) = s_x^{-n}c\cdot h\left(n\right).$$

Applying the above in the codelength used in Definition 2.26 we arrive at

$$\begin{aligned} L_{NML}^{\mathcal{M}_{\mu,\sigma}}(x^n) = & \ln f_{\mu=\hat{\mu},\sigma^2=\hat{\sigma}^2}(x^n) + DC(\mathcal{M}_{\mu,\sigma}) \\ = & \ n\ln s_x - \ln\left[c\cdot h\left(n\right)\right] + \ln\frac{2n^{\frac{n}{2}}\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)} + \ln\left[c\cdot h(n)\right] \\ = & \ n\ln s_x + \ln\frac{2n^{\frac{n}{2}}\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)}. \end{aligned}$$

□

So, according to Corrolary 3.21 spherical distributions cannot be distinguished in terms of description length. More consequences will be explored in Chapter 5.

Another parallel of Corrolary 3.21 can be established with the classical result that the task of estimating the degrees of freedom ν for Student-T distribution using an uncorrelated sample is not possible.

As noted in [Zellner, 1976], in uncorrelated samples the log-likelihood of a Student-T distribution is an increasing function of ν, thus no maximum can be achieved. Corrolary 3.21 establishes an even stronger result, that for any spherical distribution, even if an additional parameter can be estimated,

- it will necessarily be the one with the smallest complexity; and also
- cannot be a basis for inference.

Conversely, since the log-likelihood increases, and the SC criterion description length remains the same, then the complexity should decrease as a function of ν. This indeed can be seen on Figure 3.1.

3.3.1. Examples

In this section Theorem 3.19 will be used to prove that the distribution complexity of several distributions:

$$\exp DC_n\left(\mathcal{M}\right) = \begin{cases} \frac{2\left(\frac{n}{2}\right)^{\frac{n}{2}}e^{-\frac{n}{2}}}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)} & \text{for the Gaussian distribution;} \\ \frac{2n^{\frac{n}{2}}\Gamma\left(\frac{n+\nu}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)\nu^{\frac{n}{2}}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{n}{\nu}\right)^{-\frac{n+\nu}{2}} & \text{for the Student-T with } \nu \text{ df;} \\ \frac{n^n\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)\Gamma(n)}e^{-n} & \text{for the Laplace distribution.} \end{cases}$$

The three examples above are plotted on Figure 3.1.

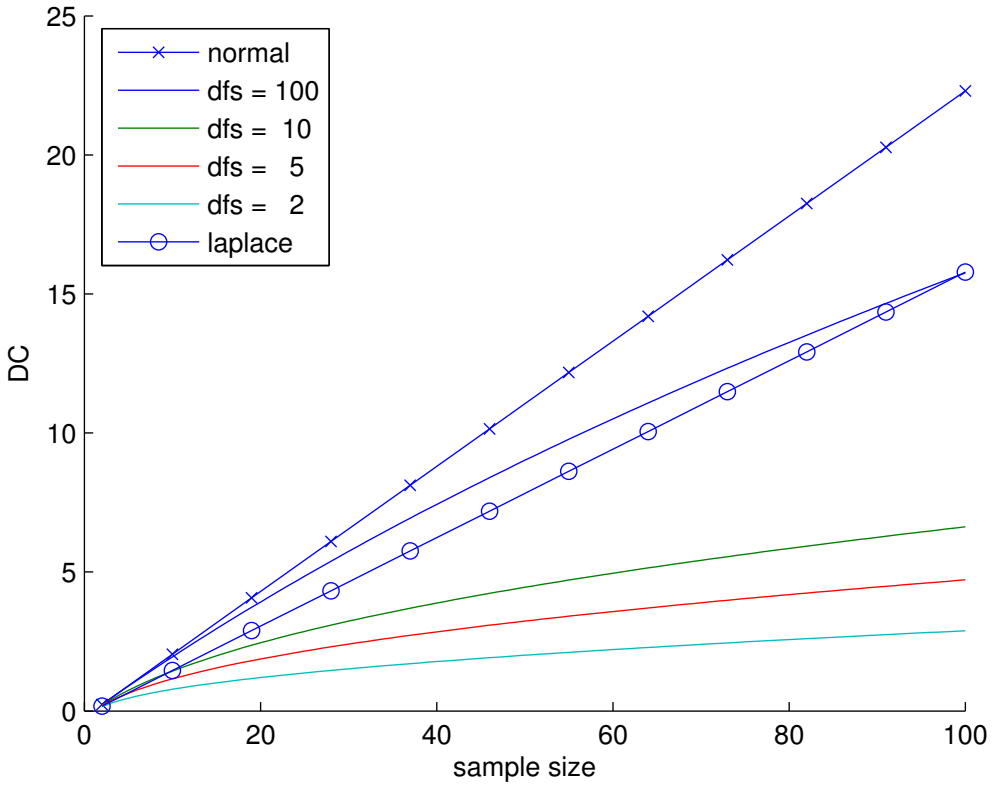

Figure 3.1.: The distribution complexity of various spherical distributions vs. sample size.

Example 3.22 (Student-T)**.** Using Definition 3.2 we can see that

$$c\sigma^{-n}K(w) = \frac{\Gamma\left(\frac{n+\nu_0}{2}\right)}{\Gamma\left(\frac{\nu_0}{2}\right)(\nu_0\pi)^{\frac{n}{2}}}s^{-n}\left(1+\frac{w}{\nu_0}\right)^{-\frac{n+\nu_0}{2}}.$$

Next we check that the regularity conditions in Definition 3.15 are satisfied:

$$\frac{\partial K}{\partial w} = -\frac{n+\nu_0}{2}\left(1+\frac{w}{\nu_0}\right)^{-\frac{n+\nu_0}{2}-1}\frac{1}{\nu_0},$$

so it exists and is negative for $\forall w > 0$. Additionally,

$$-\frac{n}{2w} = \frac{\partial \log K}{\partial w}(w) = \frac{-\frac{n+\nu_0}{2}\left(1+\frac{w}{\nu_0}\right)^{-\frac{n+\nu_0}{2}-1}\frac{1}{\nu_0}}{\left(1+\frac{w}{\nu_0}\right)^{-\frac{n+\nu_0}{2}}} = -\frac{\frac{n+\nu_0}{2\nu_0}}{1+\frac{w}{\nu_0}}$$

which can be transformed as

$$1+\frac{w}{\nu_0} = -\frac{2w}{n}\left(-\frac{n+\nu_0}{2\nu_0}\right)$$

$$w = -\nu_0 + w\left(\frac{n+\nu_0}{n}\right)$$
$$w\left(1 - \frac{n+\nu_0}{n}\right) = -v_0$$
$$w = n.$$

So we do not need to reparameterize, in which case we can apply Theorem 3.19 directly to obtain the desired decomposition.

For the Laplace distribution we will derive the results from the general form of a Laplace distribution:

$$f_{\mu,b}(x^n) = \left(\frac{1}{2b}\right)^n \exp\left\{-\frac{|x^n - \mu|}{b}\right\}.$$

To extend to spherical distribution, we will use the above univariate function, and apply a basic theorem of elliptical distributions. It can be found in [Loh et al., 1992], pp. 47.

Theorem 3.23. *Let $K(x)$ be a generating function. The symmetric univariate distribution generated by K is defined as*

$$f_{\mu,\sigma}(x) = c_1\sigma^{-1}K(\sigma^{-2}(x-\mu)^2).$$

The multivariate spherical extension of $f(x)$ is defined when $\int_0^\infty r^{n-1}K(r^2)dr < \infty$ as

$$f_{\mu,\sigma}(x^n) = c_n\sigma^{-n}K(\sigma^{-2}(x^n-\mu)^T(x^n-\mu)),$$

where the normalizing constant is defined as

$$c_n = \frac{\Gamma\left(\frac{n}{2}\right)}{2\pi^{\frac{n}{2}}\int_0^\infty r^{n-1}K(r^2)dr}.$$

The Laplace distribution is defined as the one having generator $K(w) = \exp(-\sqrt{w})$. Then

$$\int_0^\infty r^{n-1}K(r^2)dr = \int_0^\infty r^{n-1}e^{-r}dr = \Gamma(n),$$

so we can use the following definition:

Definition 3.24 (Spherical extension of Laplace distribution)**.** The Laplace distribution is defined as the one having generator $K(w) = \exp(-\sqrt{w})$:

$$f_{\mu,b}(x^n) = \frac{\Gamma\left(\frac{n}{2}\right)}{2\pi^{\frac{n}{2}}\Gamma(n)}b^{-n}\exp\left(-b^{-1}\sqrt{|(x^n-\mu)^T(x^n-\mu)|}\right).$$

We now calculate the distribution complexity of the Laplace distribution:

Example 3.25 (Laplace distribution)**.** Again, we check the regularity conditions of Definition 3.15 are satisfied:

$$\frac{\partial K}{\partial w} = \exp\left\{-\sqrt{w}\right\}\left(-\frac{1}{2\sqrt{w}}\right) \leq 0 \text{ and exists, for } \forall w > 0.$$

The last condition to be satisfied is

$$\begin{aligned}\frac{\partial \log K}{\partial w}(w) = & \frac{1}{K}\frac{\partial K}{\partial w} \\ = & \exp\left\{\sqrt{w}\right\}\exp\left\{-\sqrt{w}\right\}\left(-\frac{1}{2\sqrt{w}}\right) \\ = & -\frac{1}{2\sqrt{w}}\end{aligned}$$

The only solution for $\frac{\partial \log K}{\partial w}(w) = -\frac{n}{2w}$ then is

$$-\frac{n}{2w} = -\frac{1}{2\sqrt{w}}$$

$$w = n^2$$

so we have to normalize by changing the scale parameter to be $\sigma^2 = \frac{b^2}{n}$. Then $\tilde{K}(w) = \exp\{-\sqrt{n}\sqrt{w}\}$ and therefore

$$c\sigma^{-n}\tilde{K}(w) = \frac{\Gamma\left(\frac{n}{2}\right)}{2\pi^{\frac{n}{2}}\Gamma(n)} n^{n/2}\sigma^{-n}\exp\left\{-\sqrt{n}\sqrt{w}\right\}$$

Then we only have to apply Theorem 3.19 to the newly reparameterized $\tilde{K}$:

$$\begin{aligned}\exp DC_n(\mathcal{M}) = & \frac{2n^{\frac{n}{2}}\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)}[c \cdot K(n)] \\ = & \frac{2n^{\frac{n}{2}}\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)}\frac{\Gamma\left(\frac{n}{2}\right)}{2\pi^{\frac{n}{2}}\Gamma(n)} n^{n/2}\exp\left\{-\sqrt{n}\sqrt{n}\right\} \\ = & \frac{n^n\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)\Gamma(n)}\exp\{-n\}\end{aligned}$$

3.4. Independent scale-location families

Unlike spherical distributions, in general scale-location families do not exhibit direct connection between the sample mean and standard deviation and the MLE of the distribution parameters.

This presents a challenge when trying to apply the Stochastic Complexity (SC) criterion (Definition 2.26) similarly to Theorem 3.11 by isolating the boundary parameters R and D.

The problem is that the distribution complexity defined in Definition 3.10 does not present different distributions on an equal footing, because the parameters of the different distributions are not comparable to each other. Another way to say it is that the areas over which the constrained complexity is constructed are different.

To remedy this, this section presents more refined exploration of the structure of the distribution complexity, the main insight on which is that we can instead do a restriction on the sample mean and standard deviation, and still obtain a useful optimized method of calculation.

Definition 3.26 (Distribution complexity, general case)**.** Let $\mathcal{M}_{\mu,\sigma}$ be a scale-location family with MLE $\hat{\mu}$ and $\hat{\sigma}$. The distribution complexity of $\mathcal{M}_{\mu,\sigma}$ is defined as

$$\begin{aligned} DC_n\left(\mathcal{M}_{\mu,\sigma}\right) &= \ln \mathbb{E}_{X^n}\left[s_{X^n}\delta\left(\hat{\mu}\left(X^n\right)\left(1-\hat{\sigma}\left(X^n\right)\right)\right)\right] \\ &= \ln \int s_{x^n}\delta\left(\hat{\mu}(x^n)\right)\delta\left(1-\hat{\sigma}(x^n)\right) f(x^n)dx^n. \end{aligned} \tag{3.18}$$

Note that in the case of spherical distributions with proper parameterization $s_{y^n} = \hat{\sigma}$, and the δ-function restricts $\hat{\sigma} = 1$, so Definition 3.10 and Definition 3.26 coincide.

With the difference in definitions of the model complexity here, the following theorem applies:

Theorem 3.27 (Model complexity from DC_n, general case)**.** *Let $\mathcal{M}$ be a scale-location family and define*

$$\mathcal{B} = \{x^n : -R \leq \overline{x} \leq R, D \leq s_{x^n}\}$$

The constrained complexity can be decomposed as

$$COMP_n\left(\mathcal{M}|x^n \in \mathcal{B}\right) = \ln 2RD^{-1} + DC_n\left(\mathcal{M}\right).$$

Proof. Most of the proof follows that of Theorem 3.11. To simplify the steps, we highlight the differences between the proofs.

The first step again is to rewrite the integral using the standard density $g(x^n)$:

$$\begin{aligned} \exp COMP &= \int_{x^n\in\mathcal{B}} f\left(x^n|\mu = \hat{\mu}\left(x^n\right), \sigma = \hat{\sigma}\left(x^n\right)\right) dx^n \\ &= \int I_{\mathcal{B}}\delta\left(\mu - \hat{\mu}\left(x^n\right)\right)\delta\left(\sigma - \hat{\sigma}\left(x^n\right)\right)\sigma^{-n} f\left(\frac{x^n-\mu}{\sigma}\right) d\mu\, d\sigma\, dx^n\ . \end{aligned}$$

where $I_{\mathcal{B}}$ represents the bounds of $\overline{x}$ and s_{x^n}.

The change of variables $y^n = \frac{x^n - \mu}{\sigma}$ is made, with Jacobian determinant $|J| = \sigma^n$. An alternative form of $I_{\mathcal{B}}$ will be used to simplify the integrand:

$$\begin{aligned}
\mathcal{B}(x^n) &= \mathcal{B}(y^n, \mu, \sigma) \\
&= \left\{ y^n | \overline{y} \in \left[\frac{-R-\mu}{\sigma}, \frac{R-\mu}{\sigma} \right], s_{y^n} \in \left[\frac{D}{\sigma}, \infty \right) \right\} \\
&= \left\{ y^n | \mu \in \left[-R - \sigma\overline{y}, R - \sigma\overline{y} \right], \sigma \geq \frac{D}{s_{y^n}} \right\} .
\end{aligned}$$

Making use of lemma 3.9 and the basic properties of the δ-function from lemma B.8 we get

$$COMP_n\left(\mathcal{M}|\mathcal{B}\right) = \ln \int I_{\mathcal{B}(y^n,\mu,\sigma)} \sigma^{-2} \delta\left(\hat{\mu}\left(y^n\right)\right) \delta\left(1 - \hat{\sigma}\left(y^n\right)\right) \sigma^{-n} f\left(y^n\right) \sigma^n d\mu \, d\sigma \, dy^n.$$

The only dependence to σ and μ remaining in the integrand remains $\mathcal{B}(y^n, \mu, \sigma)\sigma^{-2}$, so we can isolate the inner integral and calculate it:

$$\begin{aligned}
\int \int I_{\mathcal{B}(y^n,\mu,\sigma)} \sigma^{-2} d\sigma d\mu &= \int_{-R-\sigma\overline{y}}^{R-\sigma\overline{y}} \int_{s_{\mathbf{y}^n}}^{\infty} \sigma^{-2} d\sigma \, d\mu \\
&= 2RD^{-1} s_{y^n}.
\end{aligned}$$

Then the complexity is expressed as

$$\begin{aligned}
COMP_n\left(\mathcal{M}|\mathcal{B}\right) &= \ln \int 2RD^{-1} s_{y^n} \delta\left(\hat{\mu}\left(y^n\right)\right) \delta\left(1 - \hat{\sigma}\left(y^n\right)\right) f\left(y^n\right) dy^n \\
&= \ln 2RD^{-1} + \ln \int s_{y^n} \delta\left(\hat{\mu}\left(y^n\right)\right) \delta\left(1 - \hat{\sigma}\left(y^n\right)\right) dF(y^n) \, . \\
&= \ln 2RD^{-1} + DC_n\left(\mathcal{M}\right).
\end{aligned}$$

□

Similarly to Theorem 3.11, Theorem 3.27 provides useful way to think and apply the SC criterion from Definition 2.26, namely by abstracting the bounds of the model complexity and allowing comparison of the competing models objectively in terms of their shape.

The application of Theorem 3.27 to numerically compute the distribution complexity is discussed in Section 4.3.

3.5. Shape parameters

In all calculations in the preceding sections the distribution family was assumed to be parameterized only by scale-location parameters.

However in real-world applications the distributions also have other parameters, which are responsible for the shape of the distribution, rather than for its scale and location. Examples include the degrees of freedom ν of a Student-T distribution, or α and β for α-Stable distribution.

Definition 3.28 (Shape parameters)**.** Let $\mathcal{M}_{\mu,\sigma,\psi}$ be a family of distributions, where μ and σ are scale-locaton parameters. ψ is called a shape parameter, if its MLE is independent of μ and σ, i.e. if $(\hat{\mu}, \hat{\sigma}, \hat{\psi})$ is the MLE, then

$$\hat{\psi}(ax^n + b) = \hat{\psi}(x^n).$$

Note that this definition is more specific in that it requires that the shape parameters estimates are independent of the scale and the location of the sample. This may require reparameterization.

Note 3.29. The converse is not necessarily true, that is, $\hat{\mu}$ and $\hat{\sigma}$ generally can and do depend on the shape parameter. This can be seen, for example, in the case of Student-T distribution, where $\mathbb{V}X = \frac{\nu}{\nu-2}\sigma^2$.

Usually we have to account for the possibility that ψ is bounded in some way, e.g. for some convex open set Ψ we have $\psi \in \Psi \cup \partial\Psi$. This provides alternative estimate, the constrained MLE:

$$\left(\tilde{\mu}, \tilde{\sigma}, \tilde{\psi}\right) = \arg_{\mu,\sigma,\psi} \max_{\psi \in \Psi \cup \partial\Psi} f_{\mu,\sigma,\psi}(x^n)$$

In this case, the following lemma gives us such a decomposition.

Lemma 3.30. *Let $\mathcal{M}_{\mu,\sigma,\psi}$ be a family of distributions and ψ to be a shape parameter constrained by the convex open set Ψ, and $\psi \in \Psi \cup \partial\Psi$, where $\partial\Psi$ is the boundary of ψ. If $\tilde{\psi}$ is the unconstrained estimator, then*

$$\begin{aligned}\exp COMP_n\left(\mathcal{M}_{\mu,\sigma,\psi}|\mathcal{A}\right) = \quad & \exp COMP\left(\mathcal{M}_{\mu,\sigma,\psi}|\mathcal{A}, \hat{\psi} \in \Psi\right) \\ & + \exp BCOMP(\mathcal{M}_{\mu,\sigma,\psi}|\mathcal{A}, \Psi)\end{aligned}$$

where the second term is called boundary complexity *and is defined as*

$$BCOMP(\mathcal{M}_{\mu,\sigma,\psi}|\mathcal{A}, \Psi) = \int_{\left\{\tilde{\psi} \neq \hat{\psi}\right\} \bigcap \mathcal{A}} f_{\tilde{\mu},\tilde{\sigma},\tilde{\psi}}(x^n) dx^n$$

Proof. First, we split x^n into those for which $\hat{\psi} = \tilde{\psi}$ and the rest. For $\hat{\psi} = \tilde{\psi}$ the boundary is not active. From linearity of the integral, and the fact that

$$\begin{aligned}\int_{\left\{\tilde{\psi}=\hat{\psi}\right\}\bigcap\mathcal{A}} f_{\hat{\mu},\hat{\sigma},\hat{\psi}}(x^n)dx^n = & \int_{\left\{\hat{\psi}\in\Psi\right\}\bigcap\mathcal{A}} f_{\hat{\mu},\hat{\sigma},\hat{\psi}}(x^n)dx^n \\ = & \exp COMP\left(\mathcal{M}_{\mu,\sigma,\psi}|\mathcal{A},\hat{\psi}\in\Psi\right).\end{aligned}$$

□

To further isolate the unneeded terms we define the constrained distribution complexity.

Definition 3.31 (Constrained Distribution Complexity)**.** Let $\mathcal{M}_{\mu,\sigma,\psi}$ be a scale-location family with shape parameter ψ. Denote by f_ψ the standard distribution for ψ, i.e. that which has $\mu = 0$ and $\sigma = 1$.

The constrained distribution complexity of $\mathcal{M}_{\mu,\sigma,\psi}$ conditional on convex open set Ψ is defined as

$$\begin{aligned}DC_n\left(\mathcal{M}_{\mu,\sigma,\psi}|\hat{\psi}\in\Psi\right) &= \ln\int_{p\in\ominus} s_{x^n}\delta\left(\hat{\mu}\right)\delta\left(1-\hat{\sigma}\right)\delta\left(p-\hat{\psi}\right)f_{\psi=p}(x^n)dx^n\,dp. \\ &= \ln\int_{p\in\ominus}\mathbb{E}\left[s_{X_p^n}\delta\left(\hat{\mu}(X_p^n)\right)\delta\left(1-\hat{\sigma}(X_p^n)\right)\delta\left(p-\hat{\psi}(X_p^n)\right)\right]dp\end{aligned}$$

Note that unlike in Definition 3.26, there is no equivalent way to express it directly as an expectation, as $f_{\hat{\psi}}$ is not a distribution function.

The above definition can be used to decompose the constrained model complexity as follows.

Lemma 3.32. *Let $\mathcal{M}_{\mu,\sigma,\psi}$ be a scale-location family with* shape parameter *ψ and define*

$$\mathcal{A} = \{-R \le \overline{x} \le R, D \le s_{x^n}\}$$

The constrained complexity can be decomposed as

$$COMP\left(\mathcal{M}_{\mu,\sigma,\psi}|\mathcal{A},\tilde{\psi}\in\Psi\right) = \ln 2RD^{-1} + DC_n\left(\mathcal{M}_{\mu,\sigma,\psi}|\tilde{\psi}\in\Psi\right).$$

Proof. Following the same steps as Theorem 3.27, and noting that by Definition 3.28 the shape parameter estimate does not depend on the scale and location. □

The last piece is the fact that the boundary component $BCOMP(\mathcal{M}_{\mu,\sigma,\psi}|\mathcal{A},\Psi)$ can also be expressed in a similar manner:

Lemma 3.33. *Let $\mathcal{M}_{\mu,\sigma,\psi}$ be a scale-location family with* shape parameter *ψ and define*

$$\mathcal{A} = \{-R \leq \overline{x} \leq R, D \leq s_{x^n}\}.$$

If $\hat{\psi}$ is the MLE constrained to the convex open set Ψ and $\tilde{\psi}$ is the unconstrained estimate, then we can show that

$$\begin{aligned} BCOMP(\mathcal{M}_{\mu,\sigma,\psi}|\mathcal{A},\Psi) = & \int_{\left\{\tilde{\psi}\neq\hat{\psi}\right\}\bigcap\mathcal{A}} f_{\tilde{\mu},\tilde{\sigma},\tilde{\psi}}(x^n)dx^n \\ = & \ 2RD^{-1}\times BDC(\mathcal{M}_{\mu,\sigma,\psi}|\Psi), \end{aligned}$$

where

$$BDC(\mathcal{M}_{\mu,\sigma,\psi}|\Psi) = \int_{\left\{\tilde{\psi}\neq\hat{\psi}\right\}} s_{x^n}\delta\left(\tilde{\mu}(x^n)\right)\delta\left(1-\tilde{\sigma}(x^n)\right)f_{\tilde{\psi}}(x^n)dx^n.$$

$BDC(\mathcal{M}_{\mu,\sigma,\psi}|\Psi)$ will be called the boundary distribution complexity.

Proof. Following the same steps as Theorem 3.27, and noting that by Definition 3.28 the shape parameter estimate does not depend on the scale and location. □

Then we can round-off with the following theorem, mirroring Theorem 3.27:

Theorem 3.34 (Model complexity with shape parameters)**.** *Let $\mathcal{M}_{\mu,\sigma,\psi}$ be a scale-location family with* shape parameter *ψ and define*

$$\mathcal{A} = \{-R \leq \overline{x} \leq R, D \leq s_{x^n}\}.$$

If $\hat{\psi}$ is the MLE constrained to the convex open set Ψ then

$$\exp COMP_n\left(\mathcal{M}_{\mu,\sigma,\psi}|\mathcal{A}\right) = 2RD^{-1}\times\left[\exp DC\left(\mathcal{M}_{\mu,\sigma,\psi}|\tilde{\psi}\in\Psi\right)+\exp BDC(\mathcal{M}_{\mu,\sigma,\psi}|\Psi)\right].$$

Proof. Directly by applying lemma 3.32 and lemma 3.33 to lemma 3.30. □

The example with Student-T distribution with non-predetermined degrees of freedom will be discussed in details in Section 4.4.

Note that, unlike in the case of pure scale-location families, we cannot discard the boundary conditions, because different distribution families have different shapes and interpretations of the shape parameter boundaries cannot be matched like it was done with the scale-location parameters.

4. Numerical Calculation

In practical applications it is seldom possible to find closed-form solutions, and for the distribution complexity we must turn to numerical methods to calculate it, except for the simple case of spherical distributions.

This chapter is structured as follows: first, we present some basic definitions, ideas and algorithms for Monte-Carlo methods in Section 4.1. In Section 4.2 we apply them to find the distribution complexity for spherical distributions from Section 3.3 and compare them with the closed-form solution derived there.

Then Section 4.3 addresses the challenges posed by the general scale-location families and provide further improvements on the formulas used for Monte Carlo integration.

The more complex case of shape parameters is described in section §Section 4.4, in the particular case of Student-T distribution, where the degrees of freedoms are free and bounded from below.

In the concluding section the numerical results are presented in form of charts and figures, useful to analyze the relations between the concepts. An application example to a real-world problem is also discussed.

4.1. Monte Carlo Integration

The basic idea of Monte Carlo integration is captured in the following definition.

Definition 4.1 (Monte-Carlo integation)**.** Let $f(x)$ be a p.d.f. of a random variable. The integral

$$\int_{-\infty}^{\infty} g(x)f(x)dx = \mathbb{E}_X g(X)$$

can be evaluated by simulating i.i.d. $X_1, \ldots, X_T \sim X$, and calculating the average provides the estimate for the integral:

$$I_T = \sum_{t=1}^{T} g(X_t) \approx \int_{-\infty}^{\infty} g(x)f(x)dx = A. \tag{4.1}$$

Note that in order for Definition 4.1 to provide valid approximations for the integral, $f(x)$ and $g(x)$ must be well-behaved. Sufficient requirement is that $\mathbb{V}g(X)$ (the

variance of $g(X)$) is finite. In such a case the variance of the estimate (4.1) is equal to $\frac{1}{T}\mathbb{V}g(X)$, and we can estimate the bounds on the estimate using Chebyshev's inequality:

$$\mathbb{P}\left(|I_T - A| \geq k\mathbb{V}g(X)\right) \leq \frac{1}{k^2T^2}.$$

There are other ways for numerical integration, but they are usually hampered by the large dimensionality of the integral in (3.8). Since we want to compute the complexity for samples, the dimension grows fast with the sample size.

4.2. Uncorrelated Samples

The direct application of Definition 4.1 does not work for the distribution complexity directly, because in (3.8) we have a generalized function under the integral:

$$DC_n\left(\mathcal{M}\right) = \mathbb{E}_{X^n}\left[\delta\left(\hat{\mu}\left(X^n\right)\left(1 - \hat{\sigma}\left(X^n\right)\right)\right)\right] = \int \delta\left(\hat{\mu}\left(x^n\right)\left(1 - \hat{\sigma}\left(x^n\right)\right)\right) f(x^n)dx^n.$$

In the notation below we will define

$$x^n = (x_1, x_2, \ldots, x_n),$$

and the subset of the first k elements will be denoted as $x^k = (x_1, x_2, \ldots, x_k)$.

Definition 4.2 (Regularity condtion for estimators)**.** Let $\mathcal{M}$ be a scale-location family $\mathcal{M}_{\mu,\sigma}$. We will say that $\mathcal{M}_{\mu,\sigma}$ satisfies a regularity condition if for any x^{n-2}, the equations

$$\hat{\mu}(x^n) = 0$$
$$\hat{\sigma}(x^n) = 1$$

conditioned on $x_{n-1} \leq x_n$ have one or zero solutions.

For the following theorem, we define the set that captures the possibility of inversion in the change of variables that we will use below.

Definition 4.3. For a scale-location family $\mathcal{M}$, the set of samples of $n-2$ elements for which we can find two elements that will give $\hat{\mu}\left(x^n\right) = 0, \hat{\sigma}\left(x^n\right) = 1$, is denoted by $\mathcal{B}^n_{\mathcal{M}}$:

$$\mathcal{B}^n_{\mathcal{M}} = \left\{x^{n-2} : \exists x_{n-1}, x_n \text{ for which } \hat{\mu}\left(x^n\right) = 0, \hat{\sigma}\left(x^n\right) = 1\right\}.$$

Thus the reason behind the regularity condition of Definition 4.2 is that for $x^{n-2} \in \mathcal{B}^n_{\mathcal{M}}$ there there is only one or two solutions.

The method of calculation is described as the following algorithm:

Algorithm 4.4 (Calculating $DC_n(\mathcal{M})$)**.** *For a spherical scale-location family $\mathcal{M}$ satisfying the regularity condition in Definition 4.2, with standard p.d.f. $f(x^n)$ we can use the following steps to obtain the distribution complexity:*

1. *Create a large simulation sample with T elements of x^{n-2}, distributed like $f(x^{n-2})$.*
2. *Do a change of variables*

$$(x_1, \ldots, x_n) \to (x_1, \ldots, x_{n-2}, m = \hat{\mu}(x^n), s = \hat{\sigma}(x^n))$$

   a) *Solve for $\hat{\mu} = 0$, $\hat{\sigma} = 1$, and obtain one solution as $x^*_{n-1} = x_{n-1}(x^{n-2}, 0, 1)$, $x^*_n = x_n(x^{n-2}, 0, 1)$.*

   b) *Record those x^{n-2} for which there is no solution.*

3. *Calculate the Jacobian determinant $|J|$ and $f(x^*_{n-1}, x^*_n | x^{n-2})$ for each element of the sample.*
4. *Calculate using $\mathcal{B}^n_{\mathcal{M}}$ from Definition 4.3 the following*

$$I(x^{n-2}) \;=\; \begin{cases} f(x^*_{n-1}, x^*_n | x^{n-2}) |J(x^{n-2}, x^*_{n-1}, x^*_n)| & \text{if } x^{n-2} \in \mathcal{B}^n_{\mathcal{M}} \\ 0 & \text{if } x^{n-2} \notin \mathcal{B}^n_{\mathcal{M}}, \end{cases}$$

5. *Obtain the approximation to distribution complexity as*

$$DC_n(\mathcal{M}) \approx \frac{2}{T} \sum_{t=1}^{T} I(x^{n-2}).$$

Proof. First, define the following change of variables

$$(x_1, \ldots, x_n) \to (x_1, \ldots, x_{n-2}, m = \hat{\mu}(x^n), s = \hat{\sigma}(x^n))$$

with Jacobian

$$|J| = \begin{vmatrix} 1 & 0 & \ldots & \frac{\partial x_1}{\partial m} & \frac{\partial x_1}{\partial s} \\ 0 & 1 & \ldots & \frac{\partial x_2}{\partial m} & \frac{\partial x_2}{\partial s} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & \frac{\partial x_{n-1}}{\partial m} & \frac{\partial x_{n-1}}{\partial s} \\ 0 & 0 & \ldots & \frac{\partial x_n}{\partial m} & \frac{\partial x_n}{\partial s} \end{vmatrix} = \begin{vmatrix} \frac{\partial x_{n-1}}{\partial m} & \frac{\partial x_{n-1}}{\partial s} \\ \frac{\partial x_n}{\partial m} & \frac{\partial x_n}{\partial s} \end{vmatrix}$$

Note that to use it in the integral, we must ensure that the function is one to one.

Since we have symmetry in the functional representations of $\hat{\mu}(x^n)$ and $\hat{\sigma}(x^n)$ when swapping x_{n-1} and x_n, we can split $\mathcal{X}^n$ in two by the hyper-plane $x_n = x_{n-1}$, and

define the area of integration as $\mathcal{A} = \{x^n : x_n \geq x_{n-1}\}$, the integral on the complement having the same value. Then by the regularity condition from Definition 4.2 we have one-to-one mapping.

We also use the formula for conditional p.d.f. to reorganize the integration:

$$f(x^n) = f(x^{n-2}) f(x_{n-1}, x_n | x^{n-2}).$$

Then we isolate the effects of the δ-function as

$$\begin{aligned} DC_n(\mathcal{M}) &= 2\int_{\mathcal{A}} \delta\left(\hat{\mu}(x^n)\left(1-\hat{\sigma}(x^n)\right)\right) f(x^{n-2}) f(x_{n-1}, x_n | x^{n-2})\, dx^n \\ &= 2\int_{\mathcal{A}} \delta\left(\hat{\mu}(x^n)\left(1-\hat{\sigma}(x^n)\right)\right) f(x_{n-1}, x_n | x^{n-2})\, dx_{n-1}\, dx_n\, dF(x^{n-2}) \\ &= 2\int \delta\left(m(1-s)\right) f(x_{n-1}, x_n | x^{n-2}) |J|\, dm\, ds\, dF(x^{n-2}) \\ &= 2\int I(y^{n-2})\, dF(x^{n-2}), \end{aligned}$$

where

$$I(x^{n-2}) = \int\int \delta\left(m(1-s)\right) f(x_{n-1}(x^{n-2}, m, s), x_n(x^{n-2}, m, s) | x^{n-2}) |J|\, dm\, ds.$$

To complete the analysis of $I(x^{n-2})$ we will use $\mathcal{B}^n_{\mathcal{M}}$ from Definition 4.3. Using the properties of δ-function from lemma B.8 we obtain

$$I(x^{n-2}) = \begin{cases} f(x^*_{n-1}, x^*_n | x^{n-2}) |J(x^{n-2}, x^*_{n-1}, x^*_n)| & \text{if } x^{n-2} \in \mathcal{B}^n_{\mathcal{M}} \\ 0 & \text{if } x^{n-2} \notin \mathcal{B}^n_{\mathcal{M}}, \end{cases}$$

where $x^*_{n-1} = x_{n-1}(x^{n-2}, 0, 1)$, $x^*_n = x_n(x^{n-2}, 0, 1)$.

This concludes the proof that the algorithm defined above calculates indeed the distribution complexity. □

The representation (3.11) gives us an optimized way to calculate the distribution complexity numerically via Monte Carlo simulations using a lot of simulated partial samples from the distribution family.

Next, we apply Algorithm 4.4 to solve the integral for the case of the model of interest - an uncorrelated Student-T distribution and Gaussian.

Since both models considered Section 3.1.1 are spherical, and with the appropriate parameterization using lemma 3.16, the MLE estimators of the parameters are the sample mean and variance:

$$\hat{\mu}(x^n) = {}^1\!/_n \sum_i x_i \text{ and } \hat{\sigma}(x^n) = {}^1\!/_n \sum_i \left(y_i - \hat{\mu}(x^n)\right)^2. \tag{4.2}$$

Lemma 4.5. *For spherical scale-location families in Algorithm 4.4,*

$$|J| = n^2 \left(\sqrt{2n - 2\left(\sum_{i=1}^{n-2} x_i^2\right) - \left(\sum_{i=1}^{n-2} x_i\right)^2} \right).$$

Proof. Using (4.2) and basic algebra. □

This approach is much better than any accept-reject method applied on (2.10) for large n, as less simulations are needed to achieve the same number of non-zero summands and consequently achieve higher accuracy.

We will then show the numerically computed values and compare them with the calculation using the analytic formula from Section 3.3.1 in Section 4.5.

4.3. General Scale-Location Families

For the more general case, covering also independent samples, the challenge in the calculation of the integral is again the dimensionality, and as remarked in [Nonchev, 2013b], the formula (4.1) successfully solves the problem. Even though the integral is high-dimensional (dimension is equal to the sample size), it can be efficiently calculated by simulating from an n-dimensional distribution.

To do the actual calculation, we can use a variant of Algorithm 4.4 that uses the general view of distribution complexity for scale-location families (Definition 3.26).

Algorithm 4.6 (Calculating $DC_n(\mathcal{M})$)**.** *For a scale-location family $\mathcal{M}$ satisfying the regularity condition in Definition 4.2, with standard p.d.f. $f(x^n)$ we can use the following steps to obtain the distribution complexity:*

1. *Create a large simulation sample of T elements of x^{n-2}, generated by $f(x^{n-2})$.*
2. *Do a change of variables*

$$(x_1, \ldots, x_n) \to (x_1, \ldots, x_{n-2}, m = \hat{\mu}(x^n), s = \hat{\sigma}(x^n)) \tag{4.3}$$

   a) *Solve for $\hat{\mu} = 0$, $\hat{\sigma} = 1$, and obtain one solution as $x^*_{n-1} = x_{n-1}(x^{n-2}, 0, 1)$, $x^*_n = x_n(x^{n-2}, 0, 1)$.*

   b) *Record those x^{n-2} for which there is no solution.*

3. *Calculate the Jacobian determinant $|J|$ and $f(x^*_{n-1}, x^*_n | x^{n-2})$ for each element of the sample.*
4. *Calculate using $\mathcal{B}^n_{\mathcal{M}}$ from Definition 4.3 the following*

$$I(x^{n-2}) = \begin{cases} s_{x^n} f(x^*_{n-1}, x^*_n | x^{n-2}) |J(x^{n-2}, x^*_{n-1}, x^*_n)| & \text{if } x^{n-2} \in \mathcal{B}^n_{\mathcal{M}} \\ 0 & \text{if } x^{n-2} \notin \mathcal{B}^n_{\mathcal{M}}, \end{cases}$$

5. *Obtain the approximation to distribution complexity as*

$$DC_n(\mathcal{M}) \approx \frac{2}{T}\sum_{t=1}^{T} I(x^{n-2}).$$

The proof parallels closely that of Algorithm 4.4, so we will not repeat it here. The only difference is the definition of $I(x^{n-2})$.

There are three challenges in the application of Algorithm 4.6, all in terms of evaluating $I(x^{n-2})$:

1. Estimation of parameters is done using numerical optimization procedure, which is computationally intensive and makes x_n and x_{n-1} somewhat unstable.
2. The relatively minor change of adding s_{x^n} has very large impact on the integral for some distributions. This is because for many distributions the MLE of the variance does not align with the sample variance, thus s_{x^n} can become very large.
3. The problem of 1. is compounded by the problem of estimating the Jacobian. It is no longer possible to find a closed-form solution, and the fact that we have to inverse it in areas where it is close to zero makes the process of estimation very unstable.

These problems are tracked in the next section, for the case of independent samples.

4.3.1. Independent Samples

The formulas in the next section are calculated with the assumption that the distribution $f_{\mu,\sigma}(x^n)$, in addition to satisfying the regularity condition in Definition 4.2, also satisfies the following smoothness constraints.

Definition 4.7 (Smoothness condition)**.** We say that a function $\mu, \sigma, x^n \to f_{\mu,\sigma}(x^n)$ satisfies a smoothness condition continuous second-order derivatives up to second-order with respect to the data $(x_1, \ldots, x_n)$ and the parameters μ and σ for both $x^n \in \mathcal{X}^n$ and $(\mu, \sigma) \in \Theta$.

Using this condition, we can find a better approximation for the Jacobian.

4.3.2. Jacobian Estimation

The main problem is the behavior of $|J^*|$ near to the boundary of $\mathcal{B}^n_{\mathcal{M}}$. In those cases when $y^*_{n-1} \to y^*_n$ the Jacobian determinant becomes unbounded, i.e. $|J^*| \to \infty$. This is a significant problem if the determinant is calculated using finite differences.

The reason it becomes unbounded is seen from the following theorem.

Theorem 4.8. *Let $\mathcal{M}_{\mu,\sigma}$ be a scale-location family which satisfies Definition 4.2 and the log-likelihood of $f_{\mu,\sigma}(x^n)$ be denoted by*

$$llf(x^n|\mu,\sigma) = \log f_{\mu,\sigma}(x^n).$$

If llf satisfies Definition 4.7, then the Jacobian of the change of variables (4.3) is then expressible as

$$\begin{pmatrix} \frac{\partial x_{n-1}}{\partial m} & \frac{\partial x_{n-1}}{\partial s} \\ \frac{\partial x_n}{\partial m} & \frac{\partial x_n}{\partial s} \end{pmatrix} = - \begin{pmatrix} \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \mu^2} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \mu \partial \sigma} \\ \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \sigma \partial \mu} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \sigma^2} \end{pmatrix} \begin{pmatrix} \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-1} \partial \mu} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_n \partial \mu} \\ \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-1} \partial \sigma} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_n \partial \sigma} \end{pmatrix}^{-1}. \tag{4.4}$$

Proof. Using the regularity condition from Definition 4.2 together with Definition 4.7, we have that for given x^{n-2} from the interior of $\mathcal{B}^n_{\mathcal{M}}$, the solutions x^*_{n-1} and x^*_n can be defined as the solutions of

$$\frac{\partial llf(x^n|\mu,\sigma)}{\partial \mu} = \frac{\partial h(x^n|\mu,\sigma)}{\partial \sigma} = 0. \tag{4.5}$$

The total derivatives with respect to m and s must be also equal to zero, since (4.5) is the definition of the MLEs (assuming x^{n-2} is fixed in advance):

$$0 = \frac{d}{dm}\left[\frac{\partial h(x^n|\mu,\sigma)}{\partial \mu}\right] = \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \mu^2} + \sum_{i=n-1}^{n} \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_i \partial \mu}\frac{dx_i}{dm}.$$

$$0 = \frac{d}{ds}\left[\frac{\partial h(x^n|\mu,\sigma)}{\partial \mu}\right] = \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \mu \partial \sigma} + \sum_{i=n-1}^{n} \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_i \partial \sigma}\frac{dx_i}{ds}.$$

$$0 = \frac{d}{dm}\left[\frac{\partial h(x^n|\mu,\sigma)}{\partial \sigma}\right] = \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \sigma \partial \mu} + \sum_{i=n-1}^{n} \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_i \partial \sigma}\frac{dx_i}{dm}.$$

$$0 = \frac{d}{ds}\left[\frac{\partial h(x^n|\mu,\sigma)}{\partial \sigma}\right] = \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \sigma^2} + \sum_{i=n-1}^{n} \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_i \partial \sigma}\frac{dx_i}{ds}.$$

The above four linear equations can be rewritten in matrix form:

$$\begin{pmatrix} \frac{\partial x_{n-1}}{\partial m} & \frac{\partial x_{n-1}}{\partial s} \\ \frac{\partial x_n}{\partial m} & \frac{\partial x_n}{\partial s} \end{pmatrix} \begin{pmatrix} \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-1} \partial \mu} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_n \partial \mu} \\ \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-1} \partial \sigma} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_n \partial \sigma} \end{pmatrix} = - \begin{pmatrix} \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \mu^2} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \mu \partial \sigma} \\ \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \sigma \partial \mu} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \sigma^2} \end{pmatrix}.$$

Multiplying both sides with the inverse directly obtains (4.4). □

Thus it is possible to obtain those partial derivatives using the partial derivatives of the log-likelihood.

Also, from (4.4) we can see that

$$|J| = \begin{vmatrix} \frac{\partial x_{n-1}}{\partial m} & \frac{\partial x_{n-1}}{\partial s} \\ \frac{\partial x_n}{\partial m} & \frac{\partial x_n}{\partial s} \end{vmatrix} = - \begin{vmatrix} \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \mu^2} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \mu \partial \sigma} \\ \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \sigma \partial \mu} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \sigma^2} \end{vmatrix} \begin{vmatrix} \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-1} \partial \mu} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_n \partial \mu} \\ \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-1} \partial \sigma} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_n \partial \sigma} \end{vmatrix}^{-1}$$

and when $x_n^* \to x_{n-1}^*$ we have that

$$\frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-1} \partial \mu} = \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_n \partial \mu}$$

$$\frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-1} \partial \sigma} = \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_n \partial \sigma}$$

then we have that

$$\begin{vmatrix} \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-1} \partial \mu} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_n \partial \mu} \\ \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-1} \partial \sigma} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_n \partial \sigma} \end{vmatrix} \to 0$$

and consequently, the Jacobian determinant is unbounded.

The good news is that since (4.4) expresses the Jacobian in terms of the partial derivatives of the log-likelihood, we can use the analytic formula in case there is one, so that the calculated Jacobian determinant will be of higher precision.

The following example is provided for the Student-T distribution.

Definition 4.9 (Independent Student-T distribution)**.** A distribution is called multivariate independent Student-T distribution with fixed degrees of freedom ν_0 if its p.d.f. is of the following form:

$$f_{\mu,\sigma}^T(x^n) = c^n \sigma^{-n} \prod_{i=1}^n \left(1 + \frac{1}{\nu_0}\left(\frac{x_i - \mu}{\sigma}\right)^2\right)^{-\frac{\nu_0+1}{2}},$$

where the normalizing constant is

$$c = \frac{\Gamma\left(\frac{\nu_0+1}{2}\right)}{\Gamma\left(\frac{\nu_0}{2}\right)\sqrt{\nu_0 \pi}}.$$

The log-likelihood is expressed as

$$\log f_{\mu,\sigma}^T(x^n) = n \log c - \frac{n}{2}\log \sigma^2 - \frac{\nu_0+1}{2}\sum_{i=1}^n \log\left(1 + \frac{1}{\nu_0}\left(\frac{x_i - \mu}{\sigma}\right)^2\right).$$

Example 4.10 (Student-T distribution)**.** Let $f_{\mu,\sigma}(x^n)$ be the Student-T distribution from Definition 4.9. The necessary condition for the MLE is

$$\frac{\partial \log f^T_{\mu,\sigma}(x^n)}{\partial \mu} = \frac{\partial \log f^T_{\mu,\sigma}(x^n)}{\partial \sigma} = 0.$$

For the Student-T distribution this is reduced as

$$0 = \frac{\partial \log f^T_{\mu,\sigma}(x^n)}{\partial \mu} = -\frac{\nu_0+1}{2}\sum_{i=1}^{n}\frac{\left(1+\frac{2}{\nu_0}\left(\frac{\mu-x_i}{\sigma^2}\right)\right)}{\left(1+\frac{1}{\nu_0}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)}$$

$$0 = \frac{\partial \log f^T_{\mu,\sigma}(x^n)}{\partial \sigma} = -\frac{n}{2}\frac{2\sigma}{\sigma^2} - \frac{\nu_0+1}{2}\sum_{i=1}^{n}\frac{\left(1+\frac{2}{\nu_0\sigma}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)}{\left(1+\frac{1}{\nu_0}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)}.$$

Because we want $\mu = 0$ and $\sigma = 1$ to be the solutions, the above deduce to

$$0 = \frac{\partial \log f^T_{\mu,\sigma}(x^n)}{\partial \mu} = -\frac{\nu_0+1}{2}\sum_{i=1}^{n}\left(1+\frac{1}{\nu_0}x_i^2\right)^{-1}\left(1-\frac{2}{\nu_0}x_i\right) \tag{4.6}$$

$$0 = \frac{\partial \log f^T_{\mu,\sigma}(x^n)}{\partial \sigma} = -n - \frac{\nu_0+1}{2}\sum_{i=1}^{n}\left(1+\frac{1}{\nu_0}x_i^2\right)^{-1}\left(1+\frac{2}{\nu_0}x_i^2\right). \tag{4.7}$$

Now, let us fix x^{n-2}. To solve the above equations, we have to find suitable x_{n-1} and x_n. The right side of second equation however increases when x_n increases or decrease, and depending on x^{n-2} it may not have a solution. On the boundary, $x_{n-1} \to x_n$, which when substituted into the Jacobian

$$\begin{vmatrix} \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-1}\partial \mu} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_n \partial \mu} \\ \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-1}\partial \sigma} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_n \partial \sigma} \end{vmatrix} \to 0 \implies |J| \to \infty.$$

Moreover, from (4.6) and (4.7) we have a very quick way to filter a lot of the samples x^{n-2} for which is impossible to find x_n and x_{n-1} that solve for the MLEs to be $\hat{\mu} = 0$ and $\hat{\sigma} = 1$.

To do that, note that only the last two summands in each case are functions of x_n and x_{n-1} and all summands in (4.7) are strictly positive. Thus (4.7) is possible only when

$$n > -\frac{\nu_0+1}{2}\sum_{i=1}^{n-2}\left(1+\frac{1}{\nu_0}x_i^2\right)^{-1}\left(1+\frac{2}{\nu_0}x_i^2\right).$$

This allows us to skip the calculation of the Jacobian for those x^{n-2} that do not satisfy this equation, which is important when n becomes large.

Let x^{n-2} be a generated sub-sample. To solve for $\hat{\mu} = 0$, $\hat{\sigma} = 1$ do a minimization of

$$u(x_n, x_{n-1}) = \left(\frac{\partial \log f^T_{\mu,\sigma}(x^n)}{\partial \mu}\right)^2 + \left(\frac{\partial \log f^T_{\mu,\sigma}(x^n)}{\partial \sigma}\right)^2.$$

This is done using the implementation of Particle Swarm Optimizer from Appendix C.

4.4. Shape Parameters

In order to apply Theorem 3.34 to a particular scale-location family with a shape parameter $\mathcal{M}_{\mu,\sigma,\psi}$, we have to calculate two integrals:

- constrained distribution complexity (Definition 3.31):

$$DC\left(\mathcal{M}_{\mu,\sigma,\psi}|\hat{\psi}\in\Psi\right)=\ln\int_{p\in\ominus}s_{x^n}\delta\left(\hat{\mu}(x^n)\right)\delta\left(1-\hat{\sigma}(x^n)\right)\delta\left(p-\hat{\psi}(x^n)\right)f_{\psi=p}(x^n)dx^n\,dp,$$

- boundary distribution complexity (lemma 3.33):

$$BDC(\mathcal{M}_{\mu,\sigma,\psi}|\Psi)=\ln\int_{\left\{\tilde{\psi}\neq\hat{\psi}\right\}}s_{x^n}\delta\left(\tilde{\mu}(x^n)\right)\delta\left(1-\tilde{\sigma}(x^n)\right)f_{\tilde{\psi}}(x^n)dx^n.$$

The algorithm to calculate $BDC(\mathcal{M}_{\mu,\sigma,\psi}|\Psi)$ depends significantly on the boundary of Ψ, so we will start with $DC\left(\mathcal{M}_{\mu,\sigma,\psi}|\hat{\psi}\in\Psi\right)$.

To do it, we can use a variant of Algorithm 4.6, for which we need a new form of regularity conditions, discussed in the next section.

4.4.1. Constrained Distribution Complexity

Definition 4.11 (Regularity condtion, shape parameters)**.** Let $\mathcal{M}$ be a scale-location family $\mathcal{M}_{\mu,\sigma,\psi}$ with shape parameter ψ . We will say that $\mathcal{M}_{\mu,\sigma,\psi}$ satisfies a *regularity condition* if for any x^{n-3}, the equations

$$\begin{aligned}\hat{\mu}(x^n)&=0\\ \hat{\sigma}(x^n)&=1\\ \hat{\psi}(x^n)&=p\end{aligned}$$

conditioned on $x_{n-2}\leq x_{n-1}\leq x_n$ have one or zero solutions.

Definition 4.12 (Smoothness condition)**.** We say that a function $\mu,\sigma,x^n\to f_{\mu,\sigma}(x^n)$ satisfies a smoothness condition continuous second-order derivatives up to second-order with respect to the data $(x_1,\ldots,x_n)$ and the parameters μ and σ for both $x^n\in\mathcal{X}^n$ and $(\mu,\sigma)\in\Theta$.

Definition 4.13. For a scale-location family $\mathcal{M}$, the set of samples of $n-3$ elements for which we can find two elements that will give $\hat{\mu}\left(x^n\right)=0,\hat{\sigma}\left(x^n\right)=1,\hat{\psi}(x^n)=p$, is denoted by $\mathcal{B}^n_{\mathcal{M}}(p)$:

$$\mathcal{B}^n_{\mathcal{M}}(p)=\left\{x^{n-3}:\exists x_{n-2},x_{n-1},x_n\text{ for which }\hat{\mu}\left(x^n\right)=0,\hat{\sigma}\left(x^n\right)=1,\hat{\psi}(x^n)=p\right\}.$$

Algorithm 4.14 (Calculating $DC\left(\mathcal{M}_{\mu,\sigma,\psi}|\tilde{\psi}\in\Psi\right)$). *For a scale-location family $\mathcal{M}$ with shape parameter ψ satisfying the regularity condition in Definition 4.11, with standard p.d.f. $f_p(x^n)$, we can use the following steps to obtain the distribution complexity:*

1. *Simulate $p\in\Psi$, possibly from a density g_p to do importance sampling.*
2. *Create a large simulation sample of T elements of x^{n-2} from $f(x^{n-2})$.*
3. *Do a change of variables*

$$(x_1,\ldots,x_n)\rightarrow\left(x_1,\ldots,x_{n-3},m=\hat{\mu}(x^n),s=\hat{\sigma}(x^n),p=\hat{\psi}(x^n)\right) \tag{4.8}$$

   a) Solve for $\hat{\mu}=0$, $\hat{\sigma}=1$, and obtain one solution as

$$\begin{aligned}x^*_{n-2}&=x_{n-2}(x^{n-3},0,1,p)\\x^*_{n-1}&=x_{n-1}(x^{n-3},0,1,p)\\x^*_{n}&=x_{n}(x^{n-3},0,1,p).\end{aligned}$$

   b) Record those x^{n-3} for which there is no solution.

4. *Calculate the Jacobian determinant $|J|$ and $f_p(x^*_{n-2},x^*_{n-1},x^*_n|x^{n-3})$ for each element of the sample.*
5. *Calculate using $\mathcal{B}^n_{\mathcal{M}}(p)$ from Definition 4.13 the following*

$$I(x^{n-3},p)\;=\;\begin{cases}s_{x^n}f_p(x^*_{n-2},\ldots,x^*_n|x^{n-3})|J(x^{n-3},x^*_{n-2},\ldots,x^*_n)| & ,\,x^{n-3}\in\mathcal{B}^n_{\mathcal{M}}\\0 & ,\,x^{n-3}\notin\mathcal{B}^n_{\mathcal{M}},\end{cases}$$

6. *Obtain the approximation of the inner-integral function as*

$$I(p)\approx\frac{6}{T}\sum_{t=1}^{T}I(x^{n-3}). \tag{4.9}$$

7. *After doing 2-6 for a large enough sample of p, calculate the approximation to distribution complexity as*

$$DC_n\left(\mathcal{M}\right)=\frac{1}{P}\sum_{i=1}^{P}\frac{I(p_i)}{g_p(p_i)}.$$

Note that in (4.9) the constant is 6, because there are six ways to reorder x_{n-2},x_{n-1} and x_n that still give a solution to $\hat{\mu}=0$, $\hat{\sigma}=1$, $\hat{\psi}=p$.

The Jacobian determinant can again be obtained using the partial derivatives of the log-likelihood, with the following theorem:

Theorem 4.15. *Let $\mathcal{M}_{\mu,\sigma,\psi}$ be a scale-location family with shape parameter ψ, which satisfies Definition 4.11 and the log-likelihood of $f_{\mu,\sigma,\psi}(x^n)$ be denoted by*

$$llf(x^n|\mu,\sigma,\psi) = \log f_{\mu,\sigma}(x^n).$$

If llf satisfies Definition 4.12, then the Jacobian of the change of variables (4.8) is then expressible as

$$\begin{pmatrix} \frac{\partial x_{n-2}}{\partial m} & \frac{\partial x_{n-2}}{\partial s} & \frac{\partial x_{n-2}}{\partial p} \\ \frac{\partial x_{n-1}}{\partial m} & \frac{\partial x_{n-1}}{\partial s} & \frac{\partial x_{n-1}}{\partial p} \\ \frac{\partial x_n}{\partial m} & \frac{\partial x_n}{\partial s} & \frac{\partial x_n}{\partial p} \end{pmatrix} = - \begin{pmatrix} \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \mu^2} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \mu \partial \sigma} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \mu \partial \psi} \\ \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \sigma \partial \mu} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \sigma^2} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \sigma \partial \psi} \\ \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \psi \partial \mu} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \psi \partial \sigma} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial \psi^2} \end{pmatrix} \times \begin{pmatrix} \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-2} \partial \mu} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-1} \partial \mu} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_n \partial \mu} \\ \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-2} \partial \sigma} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-1} \partial \sigma} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_n \partial \sigma} \\ \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-2} \partial \psi} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_{n-1} \partial \psi} & \frac{\partial^2 llf(x^n|\mu,\sigma)}{\partial x_n \partial \psi} \end{pmatrix}^{-1}. \tag{4.10}$$

Proof. Using the same steps as in Theorem 4.8. □

The actual numerical calculation using Algorithm 4.14 has not yet been achieved due to computational difficulties and is left as future work.

4.4.2. Boundary Distribution Complexity

The last term to calculate is the boundary distribution complexity. It is hard to say something meaningful about it without considering the actual boundaries, which is why the calculation is presented only for the Student-T distribution.

For Student-T distribution the shape parameter is the degrees of freedom. Letting $\psi = \nu \to \infty$ makes the Student-T distribution coincide with the Gaussian distribution, and because that reduces the diversity of the distributions (note again properties of model complexity Definition 2.25), there is no need to limit the degrees of freedom from above.

Thus $\Psi = (\nu_0, \infty)$ is a reasonable definition of the set on which the degrees of freedom are bounded.

The following result applies for the Student-T distribution.

Theorem 4.16 (Boundary Distribution Complexity for Student-T distribution)**.** *Let $\mathcal{M}^T_{\mu,\sigma,\psi}$ be the Student-T distribution family, restricted over $\nu = \psi \in \Psi = (\nu_0, \infty)$. Then*

$$BDC(\mathcal{M}^T_{\mu,\sigma,\psi}|\Psi) = DC(\mathcal{M}^{T(\nu_0)}_{\mu,\sigma}),$$

where $\mathcal{M}^{T(\nu_0)}_{\mu,\sigma}$ is the ordinary Student-T scale-location family with fixed degrees of freedom ν_0.

Proof. From the definition of $BDC(\mathcal{M}_{\mu,\sigma,\psi}|\Psi)$ we have

$$BDC(\mathcal{M}^T_{\mu,\sigma,\psi}|\Psi) = \ln \int_{\tilde{\nu}\leq\nu_0} s_{x^n}\delta\left(\hat{\mu}(x^n)\right)\delta\left(1-\hat{\sigma}(x^n)\right) f_{\hat{\nu}}(x^n)dx^n. \tag{4.11}$$

We will to analyze the function under the integral, and in particular, the relation between the unconstrained estimator $(\hat{\mu},\hat{\sigma},\hat{\psi})$ and the constrained estimator $(\tilde{\mu},\tilde{\sigma},\tilde{\psi})$.

If $\hat{\psi}=\nu_0$, $\hat{\mu}=0$ and $\hat{\sigma}=1$, then obviously $\tilde{\psi}\leq\nu_0$, because if that was not the case, then $\hat{\psi}=\tilde{\psi}$, as by construction $\tilde{\psi}$ has equal or higher likelihood than at $\hat{\psi}$.

This means that the condition $\tilde{\nu}\leq\nu_0$ in the integral (4.11) does not further restrict the range of integration, and comparing (4.11) with (3.18) completes the proof. □

Thus we can use the calculated distribution complexity of the Student-T distribution that is available in Section 4.3.

4.5. Results

4.5.1. Calculation Environment

In order to find y^*_{n-1} and y^*_nwe sill have to perform numerical optimization. This is by far the slowest part of the calculation and there is really no way around it.

To speed up the calculations, since a lot of independent optimizations are performed, an optimized massively parallel algorithm for particle swarm optimization is implemented in C++ that solves for y^*_{n-1} and y^*_n on a GPU using CUDA. It is described in more details in Appendix C.

The calculation is performed on an Intel i7-3770K with NVIDIA GTX 670 with MATLAB.

4.5.2. Uncorrelated Samples

As shown in Section 3.3, for spherical distribution scale-location families there is a closed-form solutions of the distribution complexity, at least for those that have a closed form of the density function, and also satisfy the regularity conditions.

In any case, these distributions turned out to have the same descriptive length independent of the distribution, which made them equivalent for the SC criterion from Section 2.6.3.

The results are summarized on figure Figure 4.1. The visible noise is due to the relatively low number of simulations (100000).

We can see that the distribution complexity of a the Student-T distribution is in fact lower than the complexity of the Gaussian distribution. To justify the usage of

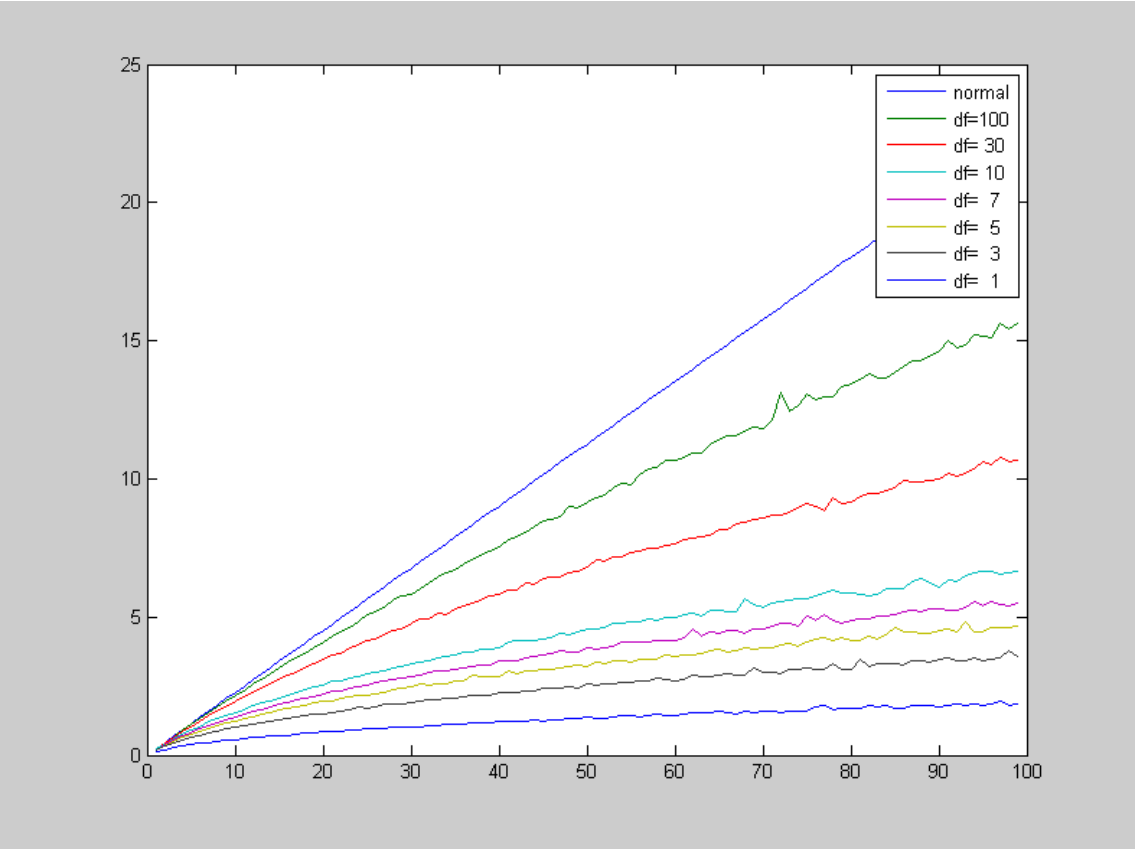

Figure 4.1.: Plot of the distribution complexity exp DC_n on the y-axis for Gaussian vs uncorrelated Student-T with various degrees of freedom, relative to the size of the n on the x-axis.

T-distribution for a sample we may have a smaller likelihood for the parameters at the MLE than for the Gaussian distribution, or in other words the T-distribution (with fixed d.f.) is actually *less complex* than the Gaussian distribution.

Another way to look at this is in terms of the model selection.

If we have a sample for which the log-likelihood in the Gaussian model is equal to the log-likelihood in the Student-T model, the MDL principle suggests that we should choose the Student-T model as more parsimonious because it is less complex.

Conversely, a sample that fits equally well the two models will have higher log-likelihood for the Gaussian model, as the Gaussian model is more complex and has a higher chance of fitting noise in the sample.

For illustrative purposes, a classical result is that for the multivariate uncorrelated Student-T distribution there is no MLE for μ, σ and the degrees of freedom ν simultaneously. This is usually interpreted through the fact that the derivative of the likelihood is positive for all ν, however it can also be considered by fact that the model complexity increases with ν and the log-likelihood will have bias toward

choosing the more complex model.

Accounting for the model complexity in the SC criterion from Definition 2.26 will also fail to choose a model, but for a different reason - because of the indifference to the actual distribution, as long as it is spherical.

4.5.3. Independent Samples

With the optimization ideas from Section 4.3 the distribution complexity is presented for several degrees of freedom for the independent Student-T distribution and compared to that of the Gaussian distribution.

Figure 4.2 summarize and compare their distribution complexity.

There is some visible noise, which is due to simulation effects. The calculations took several days on the calculation environment from Section 4.5.1

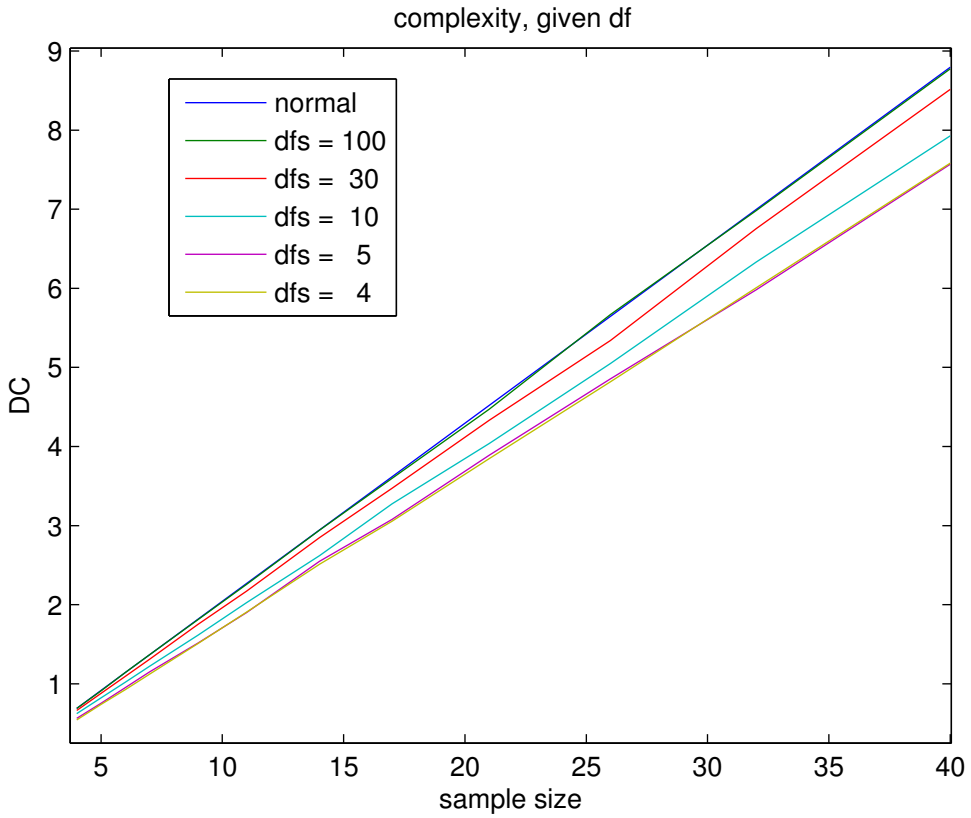

Figure 4.2.: The distribution complexity exp $DC_n(\mathcal{M})$ of independent distributions shown for different degrees of freedom.

It seems that the distribution complexity for the Student-T distribution is smaller the fatter the tails get, just like in the uncorrelated samples case in Section 4.5.2, introduced in [Nonchev, 2013a].

Unlike the uncorrelated Student-T however, there is no curvature caused by the dependence of the univariate distributions, which is very pronounced for some spherical distributions as seen on Figure 3.1.

On Figure 4.3 the DC is shown as delta compared to the Gaussian case, so that the differences can be more readily seen.

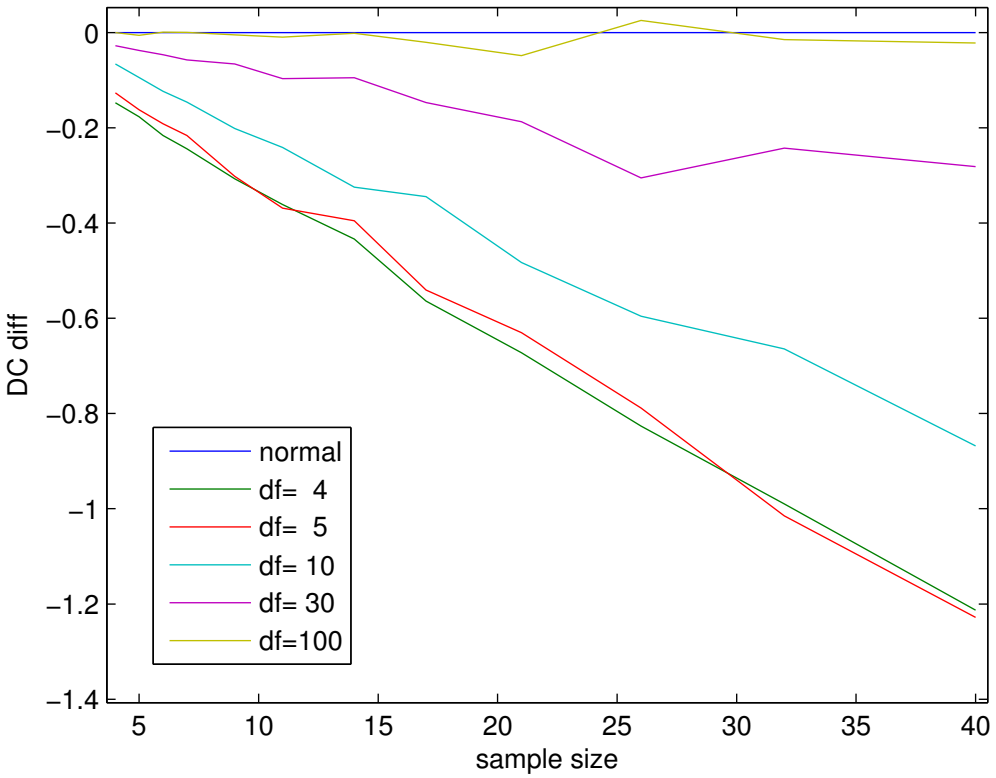

Figure 4.3.: The difference between the $\exp DC_n$ of Gaussian distribution vs $\exp DC_n$ of Student-T distribution.

The same pattern can be seen on Figure 4.4, but with the different colors now signifying the different degrees of freedom. Note that the differences, for a given sample size, of distribution complexity, are not very significant. This actually justifies criteria that discriminate between, say, Gaussian and Student-T distribution to use just the log-likelihood of the sample, because the small differences translate into small influence when applying the Stochastic Complexity criterion in Section 4.6.

The numerical results mentioned above show that with the approach outlined in this paper the numerical integration does enable us to compute a table of the distribution complexity values of Student-T distribution for a fixed degrees of freedom and size of sample up to 40 in a feasible amount of time.

The data above can be also found in Table 4.1.

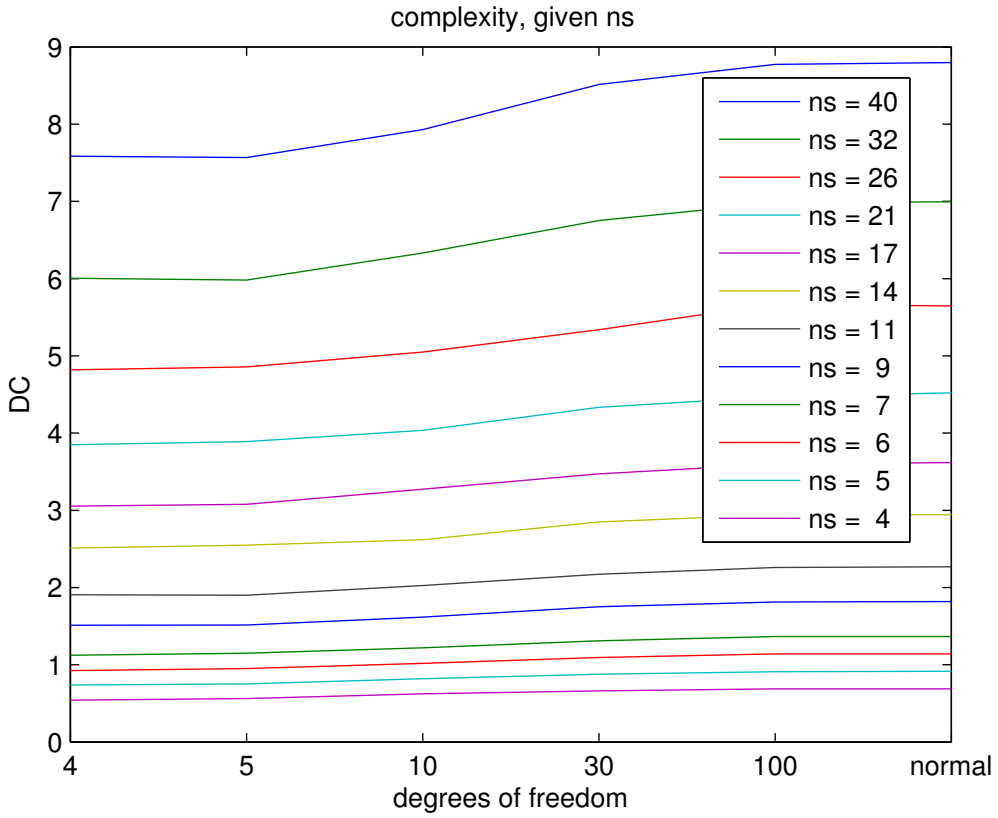

Figure 4.4.: Distribution complexity $\exp DC_n(\mathcal{M})$ of independent Student-T distribution shown for different sample sizes.

We are interested in a diversity of sample sizes. Since Table 4.1 provides values for sample sizes up to 40 elements, we have to extrapolate it. To do that we have to find a simple relationship between the values.

During the calculation of the Monte-Carlo integral a standard deviation of the values in Table 4.1 was calculated, so we do a weighted linear regression, where the explanatory variable is the sample size. The resulting values are provided in Table 4.2.

4.6. Application Example

This section discusses a simulation example to showcase the advantages of using the Stochastic Complexity criterion.

The alternative models to be used are all scale-location families, encapsulated as five hypotheses:[1]

[1] 4 degrees of freedom were dropped, because they produce very similar samples for small sample sizes.

	$\nu = 4$	$\nu = 5$	$\nu = 10$	$\nu = 30$	$\nu = 100$	$\mathcal{N}$
4	0.54	0.56	0.62	0.66	0.69	0.69
5	0.74	0.75	0.81	0.88	0.91	0.92
6	0.92	0.95	1.02	1.11	1.15	1.14
7	1.13	1.14	1.22	1.32	1.37	1.37
9	1.50	1.51	1.61	1.75	1.82	1.82
11	1.92	1.90	2.03	2.18	2.27	2.27
14	2.49	2.54	2.64	2.84	2.95	2.94
17	3.06	3.11	3.29	3.48	3.62	3.62
21	3.90	3.90	4.04	4.34	4.49	4.52
26	4.85	4.89	5.10	5.39	5.72	5.65
32	6.05	6.02	6.32	6.82	7.01	7.00
40	7.62	7.62	7.95	8.45	8.96	8.80

Table 4.1.: Distribution complexity $\exp DC_n(\mathcal{M})$ of independent Student-T distribution with ν degrees of freedom and Gaussian distribution, for sample sizes ranging from 4 to 40.

1. Independent Student-T distribution with 5 degrees of freedom;
2. Independent Student-T distribution with 10 degrees of freedom;
3. Independent Student-T distribution with 30 degrees of freedom;
4. Independent Student-T distribution with 100 degrees of freedom;
5. Independent Gaussian distribution.

All distributions above do not have a free parameter other than the scale and location, so ν is not used as a shape parameter in the sense of Section 3.5.

First some quick notes on the model selection criteria mentioned in Chapter 1 and Chapter 2:

- Since AIC (Example 1.33) and BIC (Example 1.34) only use the number of arguments as a penalty, they cannot discriminate between the Gaussian distribution and any of the Student-T distributions. Moreover, their derivation uses asymptotic considerations, which can be misleading (irrelevant) for small samples.
- Bayes factors (Example 1.32) can provide only relative evidence supporting each hypothesis.
- Stochastic Complexity criterion (Definition 2.26) fully incorporates the information can allow us to discriminate between the hypotheses with a criterion derived unambiguously from information-theoretic considerations.

	4	5	10	30	100	normal
4	0.60	0.59	0.63	0.66	0.67	0.69
5	0.79	0.79	0.84	0.88	0.90	0.92
6	0.99	0.98	1.05	1.10	1.12	1.14
7	1.19	1.18	1.26	1.32	1.34	1.37
9	1.59	1.57	1.68	1.76	1.79	1.82
11	1.99	1.97	2.09	2.20	2.24	2.27
14	2.58	2.56	2.72	2.85	2.91	2.94
17	3.18	3.15	3.35	3.51	3.58	3.62
21	3.97	3.93	4.19	4.39	4.48	4.52
26	4.97	4.92	5.24	5.49	5.60	5.65
32	6.16	6.10	6.49	6.81	6.94	7.00
40	7.75	7.67	8.17	8.56	8.73	8.80
50	9.73	9.64	10.26	10.76	10.97	11.05
63	12.32	12.20	12.98	13.61	13.88	13.97
78	15.30	15.15	16.12	16.91	17.24	17.35
98	19.27	19.08	20.31	21.30	21.72	21.85
122	24.04	23.81	25.34	26.57	27.10	27.25
244	48.48	48.01	51.10	53.58	54.64	54.96
366	72.92	72.21	76.86	80.59	82.19	82.67

Table 4.2.: Smoothed and extrapolated values of the distribution complexity $\exp DC_n(\mathcal{M})$ of independent Student-T distribution and Gaussian distribution.

4.6.1. Simulations experiment

The experiment takes the following form:

1. A number of samples is generated from each of the hypotheses.
2. For each sample, the best-fitting hypothesis is found using two criteria:

   a) Calculating SC using Definition 3.13:

   $$SC(\mathcal{M}_i) = -\log P_{\mathcal{M}_i}(y^n|\hat{\theta}(y^n)) + DC(\mathcal{M}_i).$$

   b) Just using the description length of the data $-\log P_{\mathcal{M}_i}(y^n|\hat{\theta}(y^n))$.

3. Find the best model in each case.

(a) Sample size = 21

	5	10	30	100	$\mathcal{N}$	5	10	30	100	$\mathcal{N}$
5	46.8%	9.1%	4.8%	1.9%	37.4%	5.9%	0.0%	0.6%	0.5%	-6.9%
10	32.1%	9.6%	5.5%	2.1%	50.6%	5.3%	0.7%	1.0%	0.5%	-7.5%
30	22.2%	9.0%	5.6%	2.3%	60.9%	4.6%	1.2%	1.4%	0.8%	-8.0%
100	19.6%	8.7%	5.5%	2.2%	64.0%	4.5%	1.2%	1.3%	0.7%	-7.7%
$\mathcal{N}$	18.4%	8.7%	5.4%	2.2%	65.3%	4.3%	1.5%	1.2%	0.7%	-7.7%

(b) Sample size = 50

	5	10	30	100	$\mathcal{N}$	5	10	30	100	$\mathcal{N}$
5	55.6%	17.0%	7.5%	1.6%	18.4%	3.2%	-0.5%	0.1%	-0.7%	-2.1%
10	29.8%	19.7%	10.5%	2.8%	37.2%	3.2%	0.4%	0.5%	-0.8%	-3.3%
30	14.9%	16.5%	11.3%	3.1%	54.2%	2.4%	1.1%	0.9%	-0.8%	-3.6%
100	10.7%	14.8%	10.4%	3.2%	61.0%	1.8%	1.6%	0.8%	-0.7%	-3.5%
$\mathcal{N}$	9.1%	13.8%	10.5%	2.9%	63.8%	1.7%	1.5%	1.2%	-1.0%	-3.4%

(c) Sample size = 122

	5	10	30	100	$\mathcal{N}$	5	10	30	100	$\mathcal{N}$
5	66.7%	22.2%	5.8%	1.2%	4.1%	1.9%	-0.9%	-0.3%	-0.3%	-0.3%
10	24.4%	34.4%	16.4%	4.3%	20.5%	1.8%	0.4%	-0.3%	-0.8%	-1.2%
30	5.7%	22.7%	19.4%	6.5%	45.7%	0.7%	1.3%	0.6%	-0.9%	-1.8%
100	2.7%	16.2%	17.8%	6.5%	56.8%	0.4%	1.3%	0.9%	-0.8%	-1.8%
$\mathcal{N}$	2.0%	13.5%	16.6%	6.4%	61.5%	0.3%	1.3%	0.9%	-0.7%	-1.8%

(d) Sample size = 244

	5	10	30	100	$\mathcal{N}$	5	10	30	100	$\mathcal{N}$
5	76.1%	21.0%	2.2%	0.3%	0.4%	0.9%	-0.6%	-0.3%	0.0%	0.0%
10	18.3%	49.9%	18.9%	4.3%	8.5%	1.0%	0.2%	-0.3%	-0.4%	-0.4%
30	1.4%	22.5%	26.6%	11.2%	38.3%	0.2%	1.2%	0.6%	-1.0%	-1.1%
100	0.3%	12.0%	23.7%	10.6%	53.4%	0.0%	0.9%	0.7%	-0.3%	-1.4%
$\mathcal{N}$	0.1%	8.4%	20.1%	10.3%	61.1%	0.0%	0.7%	1.1%	-0.4%	-1.4%

(e) Sample size = 366

	5	10	30	100	$\mathcal{N}$	5	10	30	100	$\mathcal{N}$
5	81.8%	17.5%	0.7%	0.0%	0.0%	1.0%	-0.9%	-0.1%	0.0%	0.0%
10	13.7%	59.3%	19.9%	3.2%	4.0%	0.8%	0.2%	-0.6%	-0.2%	-0.3%
30	0.3%	19.9%	35.1%	12.0%	32.6%	0.0%	0.8%	0.6%	-0.4%	-1.1%
100	0.0%	7.8%	27.3%	13.7%	51.2%	0.0%	0.5%	0.8%	-0.1%	-1.2%
$\mathcal{N}$	0.0%	5.1%	22.5%	13.0%	59.5%	0.0%	0.4%	1.0%	-0.3%	-1.1%

Table 4.3.: Simulation experiment results. Each rows contains the results for sample generated with the given degrees of freedom. Left part of the tables represent percentage identified by the Stochastic Complexity criterion as coming from the given degrees of freedom. Right part represents the increase attributed to the use of distribution complexity.

The results are summarized in Table 4.3, covering small sample sizes from 21 to medium samples of 366 elements.

First, analyzing the right tables we can see that the effects of the use of distribution complexity is first and foremost in the penalization of the Gaussian distribution. This is to be expected, as it has the largest complexity.

Furthermore, it is evident that the largest effect is seen on very small sample sizes, where in the extreme case of 21 samples and 5 degrees of freedom, from the 46.8% properly identified as having 5 degrees of freedom, about 5.9% percentage points, or 14% of the 46.7%, are properly recognized *because* of the penalization provided by distribution complexity. This is a significant improvement, testifying to the power of the method.

For larger samples, the effects of the distribution complexity diminishes. In our case for 244 or more elements it is so low, that in effect we are doing a Bayes factor selection, because the significant effects of the log-likelihood dominates.

Other regularities can also be seen, for example Student-T with 30 or more degrees of freedom is very hard to distinguish from the Gaussian distribution, even for larger sample sizes. This effect becomes even stronger for small sample sizes, where almost all of the samples are attributed to either to Student-T with $\nu = 5$ or to the Gaussian model.

This is expected, because in small sample sizes it is almost impossible to observe rare events, and they are the ones that determine ν the most. In those cases a single large deviation causes us to find the fattest-possible tails allowed by the hypothesis, while if no such is observed, the very complex Gaussian model fits whatever noise is present.

In conclusion, the theory behind the stochastic complexity criterion is sound, and the simulations example summarized in Table 4.3 provides evidence of the usefulness of the distribution complexity in practice. This will be applied to an actual data sample in the next section.

4.6.2. Modelling stock returns

The analysis consists of doing model selection to find the distribution of the sample from the set of alternatives discussed in Section 4.6.1. It is performed for each Friday from 1950 to 2014 (about 3300 weeks) using a rolling window. For each evaluation date, analysis using several possible values of the time window are performed:

- half-year (26 weeks);
- one year (52 weeks);
- two years (104 weeks).

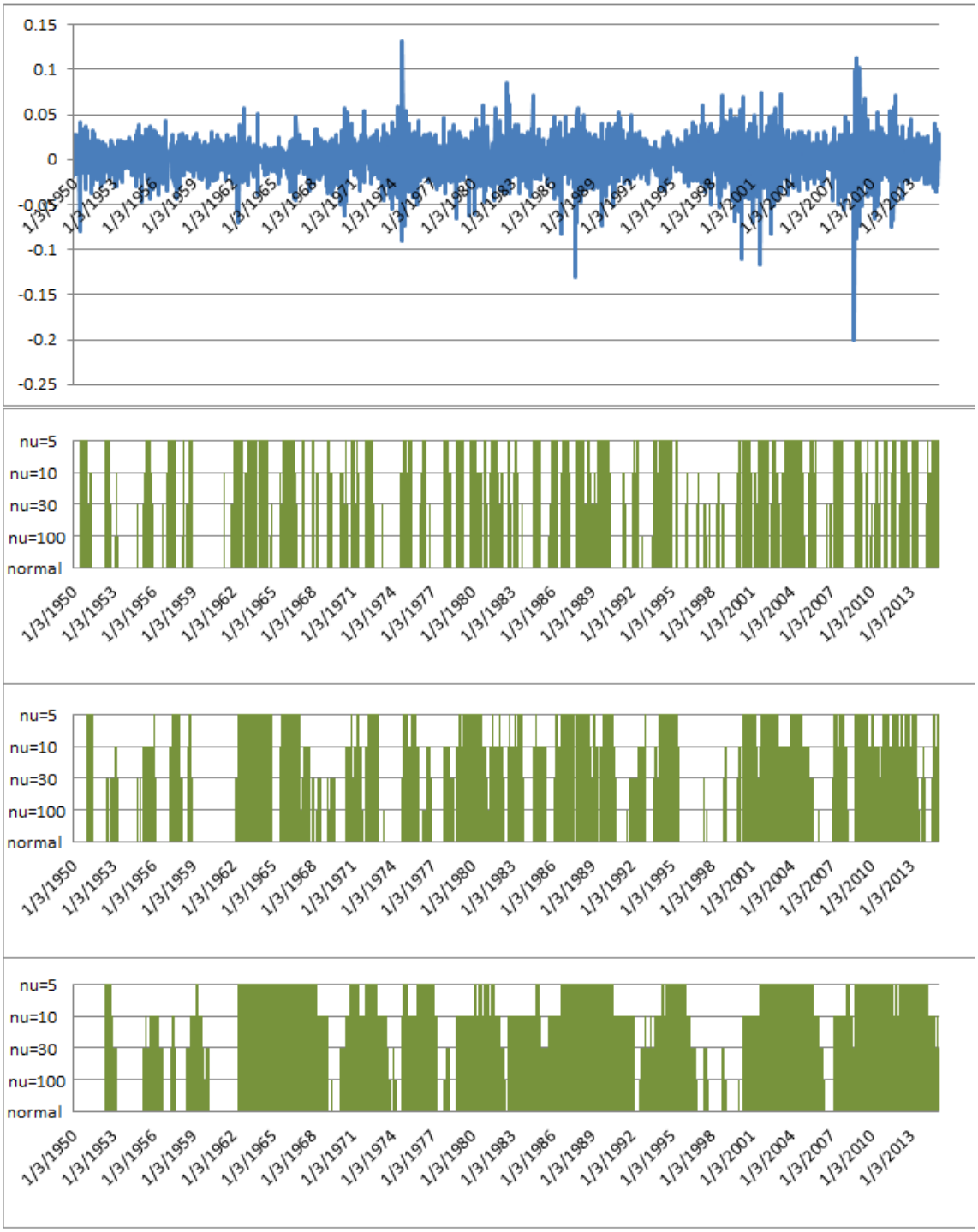

Figure 4.5.: Top, S&P 500 weekly returns from 1950 to 2014. Bottom three charts show the chosen value of the degrees of freedom from the given set, using a half-year (26 weeks), year (52 weeks) or two years worth of data.

The goal of the experiment is to pinpoint the effects of the MDL criterion, in particular the improvement over using a simple likelihood ratio test.

A basic overview of the results Figure 4.5 is shown a case study of the S&P 500 index of US stocks.[2] It is evident that the tails of the distribution are quite fat, so with the largest time window of two years, almost all of the time periods are estimated to have very fat tails. The shorter time windows show much more instability, because the extreme observations are few and far between.

In light of the discussion in the previous section we can see how when using a few data points, the best-fitting model is unstable, reverting rapidly between normal and very fat-tailed model as soon as an extreme observation comes into or exits the time window. Extending the time windows significantly stabilizes the estimated best model.

This effect can be seen as the evidence that only the Student-T model with lowest degrees of freedom is sufficiently fat-tailed to incorporate the fatter tails.

The important result in this application example is available on Table 4.4. The table shows the percentage of evaluation dates on which there is a difference between the naive log-likelihood estimator and the stochastic complexity criterion, showcasing the effects of the complexity penalization.

	half year	year	two years
difference	15.44%	12.70%	7.07%
lower df	14.52%	12.28%	6.44%
higher df	0.92%	0.42%	0.63%

Table 4.4.: Comparison between the stochastic complexity criterion (SC) and the naive log-likelihood. Table shows percentage of evaluation dates with difference in estimated degrees of freedom. Lower degrees of freedom means fatter tails are estimated when using the SC criterion.

For short time windows the effect of the distribution complexity is very large, changing the estimated degrees of freedom for 15% of the evaluation dates. Even with relative abundance of data (two years), the penalization still impacts about 7% of the evaluation dates. This means that there is a meaningful positive effect from the penalization term.

Also note that from Table 4.4 the penalization works in making the estimated distribution more heavy-tailed. This also shows the value of the penalization, since from Figure 4.5 we can see post-factum that the distribution is in fact heavy tailed, and

[2]Note that to apply DC as currently available and computed in this dissertation (using i.i.d. sample of a symmetric Student-T distribution), we have to ignore the commonly known effects such as asymmetry of stock returns and volatility clustering. This is why weekly data is used, because it exhibits them to a lesser extent than, say, daily or intraday data.

the instability of the degrees of freedom estimates for smaller time windows is due to the relative rarity of fat-tailed events in the time series.

We can conclude that the value of the distribution complexity in such cases is given by the robust principles on which this method of model selection is built, which allows us to justify its use for real world problems.

5. Conclusion

The aim of this dissertation was to solve the problem of selecting the distribution of a sample based on information theory, and more precisely the MDL principle. Since the Normalized Maximum Likelihood model (Section 2.6.2) is the optimal universal model, the natural aim was to find such a model for the task at hand.

The major obstacle when introducing the NML-based selection method for the task at hand is the fact that the model complexity is infinite (Section 2.6.4). This is a standard situation for model complexity, but problematic for our purposes.

Usually the problem is sidestepped by either using a different universal model, or doing a renormalization. These classical approaches leave a lot to be desired, especially in terms of the philosophy of the solution, which we want to preserve: the invariance with respect to parameters.

These basic settings and ideas are known in the area, and were described in Chapter 2 and can also be found in rather overwhelming details in [Grunwald, 2007]. The next section will cover the original research that was conducted in order to find a better solution to the problem of model selection.

The contributions of this dissertation are as follows:

- the introduction of distribution complexity (DC).
- practical numerical computation mechanisms to calculate the DC for a variety of distributions.
- a targeted optimizer implementation using Particle Swarm Optimization.
- calculation of DC for uncorrelated and independent samples for Student-T distribution.
- provided two application examples of the calculated distribution complexity - one with simulated samples, and one with real data sample

They are described in more detail in the next section.

5.1. Original Research

This section describes the research that was conducted and the new results obtained for this dissertation.

The problem with infinite model complexity led to the introduction of the quantity called distribution complexity (Chapter 3), which is in fact connected to the constrained complexity of the model, but also allows us to isolate the effects of the constraint's boundaries. For scale-location families, the boundaries on the scale and the location can be made equivalent for all models considered (scale-location families), which allows us to ignore them when doing model selection using the Stochastic Complexity criterion.

The first important theoretical result came from the application of this approach to spherical families (Section 3.3). These distributions do not exhibit independence for other than the multivariate Gaussian distribution, but the univariate distributions are not correlated. This would allow us to model identically distributed uncorrelated samples and to discern which distribution a sample comes from.

It turns out that there is an analytic formula for the model complexity for suitably regular distributions, and it is expressible using the p.d.f.. Unfortunately that formula, when combined with the log-likelihood of the fitted distribution cancels out into the same value for the description length, regardless of the distribution. Alternatively we can say that the description length of a sample using a spherical distribution does not depend on the distribution.

The next step was to extend the application of distribution complexity to independent distributions. This allows us to model i.i.d. samples, which is standard in statistics. The extension is done in Section 3.4 with the introduction of a slight modification of the distribution complexity formula[1]. No analytic formula is given for this case, and the remaining research for independent samples distributions is based on trying to numerically compute the values of the distribution complexity.

The numerical calculation was done in Section 4.3, where other more prosaic problems had to be overcome, for example the need for better approximation of the Jacobian of a variable change. To make the calculations faster, a particle swarm optimizer was implemented (Appendix C) which can solve the tricky problem of finding the non-zero parts of the function under the integral very fast.

Since most real-world problems use more complex distributions, which have more parameters than scale and location, the last part of the research was concentrated around trying to find a good way to find their model complexity, in particular to find some expression that has the properties of distribution complexity. This was explored in Section 3.5, which introduces integrals similar to the distribution complexity.

Unfortunately the form of these integrals depends a lot on the distribution in question, because further constraints may have to be raised and they will be on parameters which, unlike scale and location, do not compare between distributions. The particular case of Student-T distribution with additional parameters the degrees of freedom ν is explored in Section 4.4. The calculations in this are even more

[1]The new formula reduces to the old formula for spherical distributions.

complicated, and it yet remains to be seen if they are computationally tractable.

In conclusion we can say that the research done for this dissertation presents a compelling case that the MDL principle can be used for discrimination between marginal distributions of a sample.

5.2. Future Work

This section outlines several problems that seem to be practical and promising directions to extend the work presented in this dissertation.

For future work the following points and ideas deserve more elaboration:

1. *Numerical calculation of the distribution complexity example with shape parameters* - for independent Student-T with free degrees of freedom we have provided the formulas that allows the extension of the algorithm, but it has not yet been calculated with satisfactory precision.
2. *Levy α-Stable distributions and asymmetric Student-T distributions* - the model complexity of those distributions can provide a compelling case of an automatic procedure to find the distribution in econometric problems.
3. *Regression analysis* - the MDL has been used before in a regression setting, but the infinity problem has been sidestepped by using renormalization, which makes treatment of parameter values non-uniform. An approach based on distribution complexity may be more consistent with the MDL principle.
4. *ARMA, GARCH and other time-series models* - similar approach may be used to find more complex dependencies in the samples.
5. *Robust estimators* - The fact that the spherical distributions have complexity that offsets exactly their log-likelihood is somewhat surprising and it shows that the fitting method is actually the one responsible for the model complexity in the NML model, not the distribution itself. This could be extended for robust estimators, where a computing method for the parameters is usually described instead of a distribution. It is feasible that model complexity as defined by [Shtarkov, 1987], but with the new fitting mechanism, can be a sensible way to measure *estimation complexity*.

Acknowledgments

I dedicate this dissertation to my father, my mother and my brother. This thesis would not have been possible without their support and encouragement in all the years of my studies.

I want to express my gratitude to my PhD advisor for the inspiration and guidance through the fields of information theory and statistics, and the academia.

I am indebted to my many of my colleagues that supported me with ideas and challenges, including the department of Probability, Operations Research and Statistics for the logistic support and thoughful comments.

Good news, everyone!

A. Concepts in Information Theory

A.1. Overview

This section is a brief primer on the basic notions of Shannon's information theory. It follows closely [Cover and Thomas, 2012], while providing commentary and exploring the intricate connections between information theory, statistics and the MDL paradigm.

Another great introduction to information theory can be found in [Hamming, 1986].

A.2. Entropy

A.2.1. (Discrete) Entropy

Let X be a discrete random variable with a finite number of states. If the states are primarily investigated for their distinctiveness, rather than magnitude (e.g. they model categorical, rather than scale, variables), the following definition is useful:

Definition A.1. The values that X can obtain is denoted by $\aleph_X$. When X can take a finite number of states, $\aleph_X$ is called an *alphabet*.

There is a quantity called *entropy* that represents the amount of uncertainty, or information obtained by observation on the random variable. For discrete variables it is defined as follows:

Definition A.2 (Entropy of discrete r.v.)**.** The entropy of a discrete random variable is defined as

$$H(X) = -\sum_{x \in \aleph_X} p(x) \log p(x) = -\mathbb{E}_X \log p(X), \tag{A.1}$$

where $p(x)$ is the p.m.f. of X

Because $\lim_{a \to 0} a \log a = 0$, to make the above technically correct we will have to assume that the summand in case of $p(x) = 0$ is extended by continuity, i.e. $0 \log 0 = 0$.

The base of the logarithm is assumed to be 2. Since change of base represents a multiplication by a constant, the different bases will produce entropies that differ by a multiplicative constant.

The entropy is a measure of the uncertainty of the random variable, which is exemplified below.

Example A.3. Let X be a Bernoulli random variable with $p(1) = p$. Chart Figure A.1 illustrates the entropy when p varies.

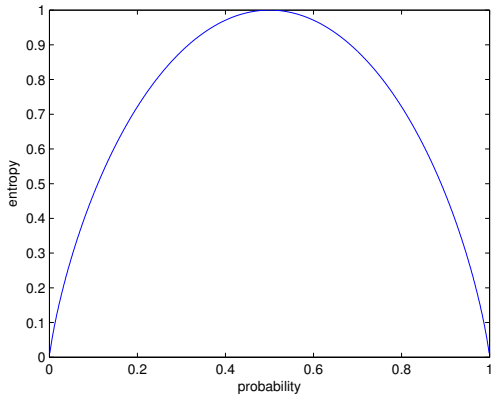

Figure A.1.: Entropy of a Bernoulli random variable for different probabilities.

Notice that when $p = 0$ or $p = 1$ the entropy $H(X) = 0$. This also is consistent with the interpretation of the entropy as uncertainty, because in that case the value of X is a.s. equal to a constant. The maximum $H(X) = 1$ is achieved when $p = \frac{1}{2}$, in which case the uncertainty of the variable is also at its maximum.

A.2.2. Differential Entropy

For absolutely continuous random variables, the entropy is infinite, so a related quantity - the differential entropy - is defined.

First we have to define the equivalent to alphabet for continuous random variables - the *support set*.

Definition A.4 (Support set)**.** For a given p.d.f. $f(x)$ the *support set* is defined as the points where the function has nonzero mass:

$$S_f = \{x \in \mathbb{R} : f(x) > 0\}$$

Definition A.5 (Differential Entropy)**.** Let X be an absolutely continuous random variable with p.d.f. $f(x)$. Its *differential entropy* is defined as

$$h(x) = -\int_{S_X} f(x) \log f(x) dx.$$

Note A.6 (bits vs. nats). So far we have not specified the base of the logarithm. Since the difference in $H(X)$ and $h(X)$ when the base of the logarithm is changed is a multiplicative constant, it does not change the properties of either. Depending on the circumstances the base is more naturally to be taken as 2 or e, in the first case we say that the entropy, and other measures, are in *bits,* and in the second case we say they are in *nats*.

Through the rest of the subsection and in the rest of the text log will refer to base-2 logarithm, unless explicitly stated. If natural logarithm should be used instead, it will be denoted by ln, and in those cases the entropy and related metrics will be measured in nats.

Note the following basic properties of differential entropy:

Theorem A.7. *Translation does not change differential entropy:*

$$h(X + c) = h(X).$$

Theorem A.8. *Scaling adds to the differential entropy*

$$h(aX) = h(X) + \log |a|.$$

For the multivariate case we have

$$h(A\mathbf{X}) = h(\mathbf{X}) + \log |\det A| \,.$$

Fact A.9. *The entropy of a multivariate Gaussian distribution with vector of means μ and covariance matrix K is*

$$h(X_1, X_2, \ldots, X_n) = h(\mathcal{N}_n(\mu, K)) = \frac{1}{2} \log(2\pi e)^n \, |K| \quad \textit{bits},$$

where $|K|$ denotes the determinant of K.

The multivariate Gaussian distribution maximizes the differential entropy over all distributions with the same covariance.

Theorem A.10. *Let the random vector $\mathbf{X} \in \mathbb{R}^n$ have zero mean and covariance $K = \mathbb{E}\mathbf{X}\mathbf{X}^t$. Then $h_e(\mathbf{X}) \leq \frac{1}{2} \ln(2\pi e)^n \, |K|$, with equality iff $\mathbf{X} \sim \mathcal{N}(0, K)$.*

Proof. Let $g(\mathbf{x})$ be the p.d.f. of $\mathbf{X}$. By definition

$$\int g(\mathbf{x}) x_i x_j d\mathbf{x} = K_{i,j} \tag{A.2}$$

for all i, j. Let ϕ_K be the density of a $\mathcal{N}(0, K)$ vector. Then $\ln \phi_K$ is a quadratic form that also satisfies (A.2). Then from the non-negativity of the relative entropy follows

$$
\begin{aligned}
0 &\leq D(g \parallel \phi_K) \\
&= \int g \ln \left(\frac{g}{\phi_K} \right) \\
&= -h(g) - \int g \ln \phi_K \\
&= -h(g) - \int \phi_K \ln \phi_K \\
&= -h(g) - h(\phi_k)
\end{aligned}
$$

where the substitution $\int g \ln \phi_K = \int \phi_K \ln \phi_K$ follows from the fact that g and ϕ_K yield the same moments of the quadratic form $\log \phi_K(\mathbf{x})$. □

Note that similar effect is observed with model complexity, where the Gaussian has large complexity than Student-T and Laplace distribution. It is an interesting (and so far open) question if that is the case for all distributions.

A.2.3. Relationship between Entropy and Differential Entropy

Discrete and differential entropy are intimately connected to each other, which can be seen using the quantization procedure from Section 1.1.3.

Theorem A.11. *If the density $f(x)$ of the random variable X is Riemann integrable, then*

$$H(X^{\Delta}) + \log \Delta \to h(f) = h(X), \text{ as } \Delta \to 0.$$

Thus the entropy of the n-bit quantization of a continuous random variable X is approximately $h(X) + n$.

Proof. From A.2 and Section A.2.2

$$
\begin{aligned}
H(X^{\Delta}) &= -\sum_i p_i \log p_i \\
&= -\sum_i p_{X^{\Delta}}(x_i) \log p_{X^{\Delta}}(x_i) \\
&= -\sum_i f(x_i^{\alpha}) \Delta \log f(x_i^{\alpha}) \Delta \\
&= -\sum_i \Delta f(x_i^{\alpha}) \log f(x_i^{\alpha}) - \log \Delta \sum_i \Delta f(x_i^{\alpha}) \\
&= -\sum_i \Delta f(x_i^{\alpha}) \log f(x_i^{\alpha}) - \log \Delta \qquad \text{(A.3)} \\
&\to h(X) - \log \Delta, \qquad \text{(A.4)}
\end{aligned}
$$

where (A.3) comes from the definition of x_i^α and (A.4) is follows whenever $f(x)$ is Riemann integrable. □

Note that this connection in some ways exemplify why the differential entropy increases with scaling of variables. If the random variable increases 2 times, then the quantization step must also increase two times in order for the set-up to be preserved.

A.2.4. Related measures

The entropy also has a host of related measures for different settings, which are described in this section.

Definition A.12 (Joint and conditional entropy, discrete case)**.** Let $(X, Y) \sim p(x, y)$. The *joint entropy* of X and Y is defined as

$$H(X,Y) = -\sum_{x\in\aleph_X, y\in\aleph_Y} p(x,y)\log p(x,y) = -\mathbb{E}_{X,Y}\log p(X,Y).$$

The *conditional entropy* is defined as

$$H(X|Y) = \sum_{x\in\aleph_X} p(y)H(X|Y=y) = -\sum_{x\in\aleph_X, y\in\aleph_Y} p(x,y)\log p(x|y) = -\mathbb{E}_{X,Y}\log p(X|Y)$$

Definition A.13 (Joint and conditional entropy, continuous case)**.** The differential entropy of a set of random variables $X_1, X_2, \ldots, X_n$ with p.d.f. f is defined as

$$h(X_1, X_2, \ldots, X_n) = -\int f(\mathbf{x}^n)\log f(\mathbf{x}^n)d\mathbf{x}^n.$$

If X, Y have a joint density function f, we can define the conditional differential entropy $h(X, Y)$ as

$$h(X|Y) = -\int f(x,y)\log f(x|y)dx\,dy = \mathbb{E}_{X,Y}\log(X|Y).$$

Conditional entropy is also related to the entropy via the following chain rules.

Lemma A.14 (Chain rule)**.**

$$H(X,Y) = H(Y) + H(X|Y)$$

$$h(X|Y) = h(X,Y) - h(Y).$$

For processes there are two definitions of entropy - entropy and entropy rate. For stationary processes they coincide.

Definition A.15 (Entropy of a process)**.** The entropy of a process X_i is defined as

$$H(X_i) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n)$$

where the limit exists.

Definition A.16 (Entropy rate of a process)**.** The entropy rate of a process X_i is defined as

$$H'(X_i) = \lim_{n\to\infty} H(X_n | X_{n-1}, \ldots, X_1)$$

where the limit exists.

Theorem A.17. *For a stationary stochastic process, the entropy and entropy rates exist and are equal:*

$$H(X_i) = H'(X_i).$$

When two or more random variables are introduced, the following related quantities are used to quantify the relation between the variables.

Definition A.18 (Relative Entropy, Kullback-Leibler divergence)**.** If $p(x)$ and $q(x)$ are two probability mass functions, the Kullback-Leibler distance between them is defined as

$$D(p \parallel q) = \sum_{x\in\aleph} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \log \frac{p(X)}{q(X)}.$$

It is finite only when $q(X) > 0$ for all x for which $p(x) > 0$.

The Kullback-Leibler divergence $D(f \parallel g)$ between two densities f and g is defined by

$$D(f \parallel g) = \int f \log \frac{f}{g} dx.$$

Note that $D(f \parallel g)$ is finite only if $S_f \subseteq S_g$.

Definition A.19. The conditional relative entropy is the average of the relative entropy between the conditional p.m.f. $p_{Y|X}$ and $q_{Y|X}$ average over the p.m.f. p_X:

$$D(p_{Y|X} \parallel q_{Y|X}) = \mathbb{E}_{p_{X,Y}} \log \frac{p_{Y,X}(Y|X)}{p_{Y,X}(Y|X)}$$

Definition A.20 (Mutual Information)**.** Let $p_{X,Y}$ be the joint p.m.f. of the random variables X and Y. The mutual information $I(X;Y)$ is defined as the relative entropy between $p_{X,Y}$ and $p_X p_y$:

$$I(X;Y) = D(p_{X,Y} \parallel p_x p_y).$$

The mutual information $I(X;Y)$ between two random variables with joint density $f_{X,Y}$ is defined as

$$I(X;Y) = D(f_{X,Y} \parallel f_x f_y) = \int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} dx\, dy.$$

This definition above works only when both X and Y are discrete or both are absolutely continuous.

There is a way to define the mutual information for any random variables, done by taking the supremum of the discrete case for the mutual information, calculated under arbitrary partition of the set of possible values for X. We use quantization of the random variables from Section 1.1.3.

Definition A.21 (Mutual Information, general case)**.** The mutual information between two random variables X and Y is given by

$$I(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I\left([X]_{\mathcal{P}}\,;[Y]_{\mathcal{Q}}\right),$$

where the supremum is over all finite partitions $\mathcal{P}$ and $\mathcal{Q}$.

Definition A.22 (Conditional Mutual Information)**.** The conditional mutual information of random variables X and Y given Z is defined by

$$I(X;Y|Z) = H(X|Z) - H(Y|Z)$$

Corollary A.23. *Properties of mutual information*

1. Information is symmetric:

   $$I(X;Y) = I(Y;X)$$

   a) Is the reduction in entropy by knowing one of the variables:

   $$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

   b) Entropy is self-information:

   $$I(X;X) = H(X)$$

Theorem A.24 (Chain rules)**.** *Let $X_1, \ldots, X_n$ be drawn according to p.m.f. $p(x_1, \ldots, x_n)$. Then*

1. *Chain rule for entropy:*

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots X_1)$$

2. *Chain rule for mutual information:*

$$I(X_1, \ldots, X_n|Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \ldots, X_1)$$

3. *Chain rule for relative entropy:*

$$D(p_{X,Y} \parallel q_{X,Y}) = D(p_X \parallel q_X) + D(p_{Y|X} \parallel q_{Y|X})$$

Theorem A.25 (Information inequality)**.** *For two p.m.f. p and q we have*

$$D(p \parallel q) \geq 0,$$

with equality iff $\forall x : p(x) = q(x)$.
For two p.d.f. f and g we have

$$D(f||g) \geq 0,$$

with equality iff $f \equiv g$.

Corollary A.26. *For two random variables X and Y,*

$$I(X; Y) \geq 0,$$

with equality iff $iX \perp Y$.

Corollary A.27.

$$D(p_{Y|X} \parallel q_{Y|X}) \geq 0,$$

with equality iff $p(x) \implies p_{Y|X}(y|x) = q_{Y|X}(y|x)$.

Corollary A.28.

$$I(X; Y|Z) \geq 0,$$

with equality iff X and Y are conditionally independent given Z.

Theorem A.29. *If $|\aleph_X| < \infty$ then*

$$H(X) \leq \log |\aleph_X| .$$

The equality is achieved when X is uniform over $\aleph_X$.

Theorem A.30 (Information can't hurt)**.** *If X and Y are discrete random variables, then*

$$H(X|Y) \leq H(X),$$

with equality iff $X \perp Y$.

If X and Y are absolutely continuous, then

$$h(X|Y) \leq h(X),$$

with equality iff $X \perp Y$.

Theorem A.31 (Independence bound of entropy)**.**

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i),$$

with equality iff X_i are independent.

Theorem A.32 (Data-processing inequality)**.** *If $X \to Y \to Z$ (i.e. form a Markov chain, see 1.5), then*

$$I(X;Y) \geq I(X;Z).$$

Corollary A.33. *In particular, if $Z = g(Y)$, then $I(X;Y) \geq I(X;g(Y))$.*

This means that no matter how we process the variable Y, the information about X that it contains cannot increase.

A.3. Asymptotic Equipartition Property

A basic notion from information theory that relates the entropy of a random variable to its compression, and that of samples derived by it, is the asymptotic equipartition property. It loosely states that for random variables with low entropy there are subsets of samples that:

1. are small, i.e. have few elements; and
2. are highly probable, with probability almost 1.

These subsets are called *typical sets* and are instrumental in the links between entropy and compression.

First they are introduced for the discrete case, and then are naturally extended for the continuous case.

A.3.1. Discrete case

Theorem A.34 (Asymptotic equipartition property (AEP))**.** *If $X_1, X_2, \ldots$ are i.i.d. with p.m.f. p_X, then*

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X) \text{ in probability.}$$

This means that the probability of events become very similar for most of the possible events. On that basis we define sets containing that majority of events:

Definition A.35. The *typical set* $A_\epsilon^{(n)}$with respect to p.m.f. p is the set of sequences $(x_1, \ldots, x_n) \in \aleph^n$ with the property that

$$2^{-n(H(X)+\epsilon)} \leq p\left(x_1, \ldots, x_n\right) \leq 2^{-n(H(X)-\epsilon)}.$$

Theorem A.36 (Properties of the typical set)**.**

1. Typical set element have log-probability about equal to the entropy: if $(x_1, \ldots, x_n) \in A_\epsilon^{(n)}$, then

$$H(X) - \varepsilon \leq -\frac{1}{n}\log p\left(x_1, \ldots, x_n\right) \leq H(X) + \varepsilon.$$

   a) $\mathbb{P}\left\{A_\varepsilon^{(n)}\right\} > 1 - \varepsilon$ for sufficiently large n.

   b) The number of elements in the typical set is about $2^{nH(X)}$:

$$(1-\varepsilon)2^{n(H(X)-\varepsilon)} \leq \left|A_\varepsilon^{(n)}\right| \leq 2^{n(H(X)+\varepsilon)}.$$

Example A.37 (Data compression algorithm)**.** There is a way to efficiently encode the outcome of an experiment $(X_1, X_2, \ldots, X_n)$ using the typical set $A_\varepsilon^{(n)}$:

1. Fix n and δ and construct $A_\varepsilon^{(n)}$ *with* $\epsilon = \frac{\delta}{n\log|\aleph_X|+1}$.

   a) Use 1 bit to encode whether the event is in the typical set.

   b) Enumerate the approximately $2^{nH(X)}$ elements of the typical set with about $nH(X)$ bits.

   c) Enumerate the remaining elements, which are about $|\aleph_X|^n$ with $n\log|\aleph_X|+1$ bits.

Example A.38. The expected length of code will be

$$\mathbb{E}L(X) = (1-\delta)nH(X) + \delta \to nH(X).$$

From the A.35 it is clear that typical sets are fairly small set, but due to their construction contain most of the probability. This observation is extended to define *high-probability sets*, which turn out to mostly coincide with typical sets.

Definition A.39 (High-probability set)**.** For each $n = 1, 2, \ldots$ let $B_\delta^{(n)} \subset |\aleph|^n$ be the smallest set with

$$\mathbb{P}\left\{B_\delta^{(n)}\right\} > 1 - \delta.$$

Theorem A.40. *Let $X_1, \ldots, X_n$ be i.i.d. with p.m.f. $p(x)$. For $\delta < 1/2$ and any $\delta' > 0$, if $\mathbb{P}\left\{B_\delta^{(n)}\right\} > 1 - \delta$, then*

$$\frac{1}{n} \log \left|B_\delta^{(n)}\right| \geq H(X) - \delta'$$

Therefore for n sufficiently large we have that the number of elements is approximately $2^{nH(X)}$, just like the typical set.

A.3.2. Continuous case

Theorem A.41. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sequence with density $f(x)$. Then*

$$-\frac{1}{n} \log f(X_1, X_2, \ldots, X_n) \to \mathbb{E}\left[-\log f(X)\right] = h(X) \text{ in probability.}$$

Definition A.42. For any $\varepsilon > 0$, the typical set $A_\epsilon^{(n)}$with respect to $f(x)$ is defined as

$$A_\epsilon^{(n)} = \left\{(x_1, \ldots, x_n) \in S^n : \left|-\frac{1}{n} \log f(x_1, x_2, \ldots, x_n) - h(X)\right| \leq \varepsilon\right\},$$

where f the set of sequences $(x_1, \ldots, x_n) \in \aleph^n$ with the property that

$$2^{-n(H(X)+\epsilon)} \leq p\left(x_1, \ldots, x_n\right) \leq 2^{-n(H(X)-\epsilon)}.$$

Definition A.43. The volume $Vol(A)$ of a set $A \subset \mathbb{R}^n$ is defined as

$$Vol(A) = \int_A dx_1 dx_2 \ldots dx_n.$$

Theorem A.44. *The typical set $A_\epsilon^{(n)}$ has the following properties:*

1. $\mathbb{P}\left\{A_\varepsilon^{(n)}\right\} > 1 - \varepsilon$ for sufficiently large n.

   a) The volume of the typical set is about $2^{nH(X)}$, i.e.

   $$(1 - \varepsilon)2^{n(h(X)-\varepsilon)} \leq Vol\left(A_\varepsilon^{(n)}\right) \leq 2^{n(h(X)+\varepsilon)},$$

   the upper bound for all n and the lower bound for n sufficiently large

Theorem A.45. *The set $A_\epsilon^{(n)}$ is the smallest volume set with $\mathbb{P}\left((X_1 \ldots, X_n) \in A_\epsilon^{(n)}\right) \geq 1 - \varepsilon$.*

A.4. Source coding

Definition A.46. A *source code* $C_{\aleph}$ for a random variable X is the mapping from its alphabet $\aleph$ to the set of finite-length $\mathcal{D}$-ary strings $\mathcal{D}^*$. $C(x)$ denote the codeword corresponding to x and $l(x)$ is the codelength.

Definition A.47. The *expected length* $L_X(C)$ of a source code $C(x)$ for a random variable X with p.m.f. $p(x)$ is given by

$$L_X(C) = \sum_{x \in \aleph_X} p(x)l(x) = \mathbb{E}l(X),$$

where $l(x)$ is the length of the codeword associated with x.

Definition A.48. A code is said to be *non-singular* if different elements of $\aleph_X$ map to different strings in $\mathcal{D}^*$; that is

$$x \neq x' \implies C(x) \neq C(x').$$

i.e. the code is reversible.

Definition A.49. The *extension* C^* of a code C is the mapping from finite-length strings of $\aleph_X$ to finite-length strings of $\mathcal{D}$, defined by

$$C^*(x_1x_2 \dots x_n) = C(x_1)C(x_2) \dots C(x_n),$$

The operation on the right hand side is concatenation of corresponding codewords.

Definition A.50. If the extension of a code is non-singular, the code is called *uniquely decodable.*

Definition A.51. A code is called a prefix code or an instantaneous code if no codeword is prefix for any other codeword.

The next property we introduce is Kraft's inequality. It is very important characterization of codes and in general justifies the restriction of uniquely decodable codes to prefix codes, since we gain nothing from using the more general property of unique decodability.

Theorem A.52 (Kraft's inequality)**.** *For any countably infinite set of codewords that form a prefix code, the codelengths satisfy*

$$\sum_i D^{-l_i} \leq 1,$$

This is called Kraft's inequality, and l_i is the length of the i-th codeword, $i \in \mathbb{N}$.

Note that in Kraft's inequality we do not concern ourselves with the actual values assigned to the codewords.

The following theorem proves that we do not gain anything from considering non-prefix codes, so we might as well enjoy the advantages that prefix codes offer. For shortness, this stronger version will be referred to as Kraft's inequality later in the test.

Theorem A.53. *[McMilan] The codeword lengths of any uniquely-decodable D-ary code must satisfy Kraft's inequality*

$$\sum D^{-l_i} \leq 1. \tag{A.5}$$

Conversely, for a given set of codeword lengths that satisfy the inequality it is possible to construct a uniquely decodable code.

If a code satisfies (A.5) with equality, it is called Kraft-tight. Intuitively, Kraft-tight codes are optimal in a sense that there is no code that has equal or better codelength for all codewords, because their existence contradicts Theorem A.53.

Theorem A.54. *The expected length of any instantaneous D-ary code for a random variable X is greater or equal to the entropy $H_D(X)$; i.e.*

$$L(C_{\aleph_{\ddot{\imath}}}) \geq H_D(X),$$

with equality iff $D^{-l(x)} = p(x)$.

Proof. The difference between the expected length and the entropy is

$$\begin{aligned} L(C_{\aleph_{\ddot{\imath}}}) - H_D(X) &= \sum p_i l_i - \sum p_i \log_D \frac{1}{p_i} \\ &= \sum p_i \log_D D^{l_i} - \sum p_i \log_D \frac{1}{p_i} \\ &= \sum p_i \log_D \frac{p_i}{D^{-l_i}} \\ &= \sum p_i \log_D \frac{p_i}{r_i} + \log_D \frac{1}{c} \\ &= D(p||r) + \log_D \frac{1}{c}, \end{aligned}$$

where $r_i = \frac{D^{-l_i}}{\sum_j D^{-l_j}}$ and $c = \sum_j D^{-l_j} \leq 1$ by Kraft's inequality. □

Definition A.55. A probability distribution is called D-adic if each of the values of $p(x)$ is equal to D^{-n} for some n.

The optimal code for D-adic distributions can very easily be obtained by matching each element of $\aleph_X$ to a leaf of a D-adic tree. For other distributions one way to find optimal code is to find the D-adic distribution that has the least relative entropy (K-L divergence), see A.18. This however is hard to compute. There is an easier approach, which gives an "almost" optimal code, the so-called *Shannon code*.

Theorem A.56 (Shannon code)**.** *Let X be a random variable. There is code with lengths assigned as*

$$l_i = \lceil -\log_D p_i \rceil \tag{A.6}$$

that satisfies the bounds

$$H_D(X) \leq L < H_D(X) + 1, \tag{A.7}$$

called Shannon code*.*

Proof. To construct the Shannon's code we create a matching code with D-adic tree. That is possible, because the lengths satisfy Kraft's inequality trivially:

$$\sum_i D^{-l_i} = \sum_i D^{-\lceil -\log_D p_i \rceil} \leq \sum_i D^{\log_D p_i} = \sum_i p_i = 1.$$

Moreover, from the definition directly follows that

$$-\log_D p_i \leq l_i < -\log_D p_i + 1$$

and weighting with p_i and summing we obtain (A.7). □

Since an optimal code is no worse than Shannon code, the following Corollary is proven:

Corollary A.57. *Let $l^*(x)$ be the optimal codeword lengths for a random variable X in a D-ary alphabet, and let L_X^* be its associated expected length. Then*

$$H_D(X) \leq L_X^* \leq H_D(X) + 1.$$

To obtain further compression, we can merge together consecutive observations of X.

Corollary A.58. *Let $X_1, \ldots, X_n$ be i.i.d. The minimum expected codeword length per symbol satisfies*

$$\frac{H(X_1, \ldots, X_n)}{n} \leq L_n^* < \frac{H(X_1, \ldots, X_n)}{n} + \frac{1}{n}.$$

Moreover, if $X_1, \ldots, X_n$ is a stationary stochastic process, then

$$L_n^* \to H\left(\{X_n\}\right),$$

where $H\left(\{X_n\}\right)$ is the entropy rate of the process.

Theorem A.59 (Wrong code)**.** *The expected length under $p_X(x)$ of the code assignment $l(x) = \left\lceil \log \frac{1}{p_Y(x)} \right\rceil$ satisfies*

$$H_D(X) + D(X \parallel Y) \leq \mathbb{E}_X l(X) \leq H_D(X) + D(X \parallel Y) + 1.$$

The following theorem shows that the Shannon code is very good even on individual-sequence level:

Theorem A.60 (Competitive optimality of the Shannon code)**.** *Let $l(x)$ be the codeword lengths associated with the Shannon code for X and let $l'(x)$ be the code with any other uniquely decodable code. Then*

$$\mathbb{P}\left(l(X) \geq l'(X) + c\right) \leq \frac{1}{2^{c-1}}$$

That is, no other code can be better than the Shannon code with high probability.

The following is the explicit scheme of the Shannon-Fano code.

Example A.61 (Shannon-Fano code)**.** Shannon-Fano codes are constructed recursively, via the following divide-and-conquer algorithm.

Let $S \subseteq \mathcal{X}$ be a set of possible outcomes. For them:

1. If the set contains only one outcome, then it has been properly encoded.
2. Sort the list by each outcome's probability.
3. Split the list into S_0, S_1 at the place which makes the sum of the probabilities on the left as close as possible to those on the right.
4. Encode the elements of S_0 recursively with steps 1-4, and add a 0 in front of their codeword. Encode the elements of S_1 recursively with steps 1-4 and add 1 in front of their codewords.

A.5. Application in Statistics

A.5.1. Sufficient statistics

This section is continuation of Section 1.1.2, expanding and refining the notion of sufficient statistic to connect it with information theory concepts.

Definition A.62. A function $T(X)$ is called *sufficient statistic* relative to the family of distributions $\{F_\theta(x)\}$ if X is independent of θ given $T(X)$ for any distribution on θ, i.e. $\theta \to T(X) \to X$ forms a Markov chain.

Definition A.63. A statistic $T(X)$ is called a *minimal sufficient statistic* relative to $\{F_\theta(x)\}$ if it is a function of every sufficient statistic, i.e.

$$\theta \to T(X) \to U(X) \to X.$$

That means that the information contained in U may reduce the uncertainty in X further than $T(X)$ would, but it is not more connected to θ than what is seen through $T(X)$.

Theorem A.64 (Fano's inequality)**.** *For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, with $P_e = \mathbb{P}\left\{X \neq \hat{X}\right\}$, we have*

$$H(P_e) + P_e \log |\aleph_X| \geq H(X|\hat{X}) \geq H(X|Y).$$

This can be weakened to

$$P_e \geq \frac{H(X|Y) - 1}{\log |\aleph_X|}$$

A.5.2. Universal Source Coding

An important concept from information theory, that is directly applicable in the MDL is the concept of universal source codes. Using such codes we are able to achieve compression without explicitly working out the exact distribution.

Definition A.65. A *fixed-rate block code* of rate R for a source $X_1, X_2, \ldots, X_n$ which has unknown distribution Q consists of two mappings: the encoder,

$$f_n : \aleph^n \to \left\{1, 2, \ldots, 2^{nR}\right\},$$

and the decoder,

$$\phi_n : \left\{1, 2, \ldots, 2^{nR}\right\} \to \aleph^n.$$

Here R is called the rate of the code. The probability of error for the code with respect to the distribution Q is

$$P_e^{(n)} = Q^n(X^n : \phi_n(f_n(X^n)) \neq X^n)$$

Definition A.66. A rate R block code for a source will be called *universal* if the functions f_n and ϕ_n do not depend on the distribution Q and if $P_e^{(n)} \to 0$ as $n \to \infty$ if $R > H(Q)$.

Theorem A.67. *There exists a sequence of $\left(2^{nR}, n\right)$ universal source codes such that $P_e^{(n)} \to 0$ for every source Q such that $H(Q) < R$.*

Proof. Fix the rate R of the code. Let $R_n = R - |\aleph| \frac{\log(n+1)}{n}$ and consider the set of sequences

$$A = \{\mathbf{x} \in \aleph^n : H(P_{\mathbf{x}}) \leq R_n\}.$$

Then

$$\begin{aligned}
|A| &= \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} |T(P)| \\
&\leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nH(P)} \\
&\leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nR_n} \\
&\leq (n+1)^{|\aleph|} 2^{nR_n} \\
&= 2^{n\left(R_n + |\aleph| \frac{\log(n+1)}{n}\right)} \\
&= 2^{nR}
\end{aligned}$$

So we can map directly each of the 2^{nR} elements of A. All we have to prove is that the scheme is universal. Let Q be the distribution of $X_1, \ldots, X_n$ and $H(Q) \leq R$. The probability of decoding error is then given by

$$\begin{aligned}
P_e^{(n)} &= 1 - Q^n(A) \\
&= \sum_{P : H(P) > R_n} Q^n(T(P)) \\
&\leq (n+1)^{|\aleph|} \max_{P : H(P) > R_n} Q^n(T(P)) \\
&\leq (n+1)^{|\aleph|} 2^{-n \min_{P : H(P) > R_n} (D(P \| Q))}.
\end{aligned}$$

Since $R_n \uparrow R$ and $H(Q) < R$, there exists n_0 such that for all $n \geq n_0$, $R_n > H(Q)$. Then for $n \geq n_0$, $\min_{P : H(P) > R_n} D(P \parallel Q) > 0$, so the probability of error $P_e^{(n)}$ converges to 0 exponentially fast when $n \to \infty$. □

A.5.3. Estimation error

Using the properties of the multivariate Gaussian distribution from Section A.2.2, one can prove that the estimation error is capped from below by approximately the volume of the typical set (see Theorem A.44). Additionally, the estimation error can achieve the boundary only in the Gaussian case.

This is presented in the following theorem.

Theorem A.68 (Estimation error and differential entropy)**.** *Let X be a random variable and $\hat{X}$ an estimator, then*

$$\mathbb{E}(X - \hat{X})^2 \geq \frac{1}{2\pi e} e^{2h(X)}.$$

Equality is achieved iff X is Gaussian and $\hat{X}$ is the mean of X.

Corollary A.69. *Given side information Y and estimator $\hat{X}(Y)$ it follows that*

$$\mathbb{E}(X - \hat{X}(Y))^2 \geq \frac{1}{2\pi e} e^{2h(X|Y)}.$$

A.5.4. Fisher Information

Let $\{f_\theta(x)\}_{\theta\in\Theta}$ denote an indexed family of densities. Θ is called the parameter set.

Definition A.70. The *score* V is a random variable defined by

$$V = \frac{\partial}{\partial\theta} \ln f_\theta(X) = \frac{\frac{\partial}{\partial\theta} f_\theta(X)}{f_\theta(X)}$$

The mean value of the score is zero, $\mathbb{E}V = 0$, hence the $\mathbb{E}V^2 = var(V)$ and the variance of the score has a special significance.

Definition A.71. The *Fisher information* $J(\theta)$ is the variance of the score:

$$J(\theta) = \mathbb{E}_\theta \left[\frac{\partial}{\partial\theta} \ln f_\theta(X) \right]^2 .$$

Let $X_1, \ldots, X_n$ be i.i.d. random variables. Then

$$V(X_1, \ldots, X_n) = \frac{\partial}{\partial\theta} \ln f_\theta(X_1, \ldots, X_n) = \sum_i \frac{\partial}{\partial\theta} \ln f_\theta(X_i) = \sum_i V(X_i)$$

and consequently

$$\begin{aligned} J_n(\theta) &= \mathbb{E}_\theta \left[\frac{\partial}{\partial\theta} \ln f_\theta(X_1, \ldots, X_n) \right]^2 \\ &= \mathbb{E}_\theta \left[\sum_i V(X_i) \right]^2 \\ &= \sum_i \mathbb{E}_\theta \left[V(X_i) \right]^2 \\ &= nJ(\theta) \end{aligned}$$

The important property of Fisher information is that it bounds the error of estimators:

Theorem A.72 (Cramér-Rao inequality)**.** *The mean-squared error of any unbiased estimator $T(X)$ of the parameter θ is lower bounded by the reciprocal of the Fisher information:*

$$var(T) \geq \frac{1}{J(\theta)}.$$

Definition A.73. An unbiased estimator T is said to be efficient if it meets the Cramér-Rao bound with equality, i.e. $var(T) = \frac{1}{J(\theta)}$.

This can be generalized to a multivariate setting, where the score is a random vector and the Fischer information matrix $J(\theta)$ is defined with elements

$$J_{i,j} = \int f_\theta(x) \frac{\partial}{\partial \theta_i} \ln f_\theta(x) \frac{\partial}{\partial \theta_j} \ln f_\theta(x) dx = \mathbb{E} V V^T.$$

In this setting the Cramér-Rao inequality can be rewritten as

$$\Sigma \geq J^{-1},$$

where Σ is the covariance matrix of a set of unbiased estimators for the parameters θ and the inequality is in the sense that the difference is a non-negative definite matrix.

Theorem A.74 (de Bruijn's identity: entropy and Fisher information)**.** *Let X be any random variable with a finite variance with a density $f(x)$. Let Z be an independent Gaussian distributed random variable with zero mean and unit variance. Then*

$$\frac{\partial}{\partial t} h_e(X + \sqrt{t}Z) = \frac{1}{2} J(X + \sqrt{t}Z),$$

where h_e is the differential entropy to base e. In particular, if the limit exists as $t \to 0$,

$$\frac{\partial}{\partial t} h_e(X + \sqrt{t}Z)\bigg|_{t=0} = \frac{1}{2} J(X).$$

A.6. Kolmogorov Complexity

The concept of Kolmogorov complexity was introduced in [Kolmogorov, 1963]. A good introduction to the topic and its applications is available in [Li and Vitanyi, 1993].

For our purposes we will introduce some concepts required to make sense of the MDL paradigm in Chapter 2.

A classic example problem used in the inspirational paper of Kolmogorov is that if you are charged with transmitting three sequences of a million symbols 0 or 1 each, e.g.

- 01010101010101010101010101010101010...
- 11011001111111011111101100111111111111...
- 10101010001110100011101000111010111110...

you can certainly do better than transmitting the whole sequence bit by bit, if you exploit regularities in the data.

The first sequence is just 01 repeated, so sending this instruction instead is a lot easier. The second has about 9 ones for each zero, so we can encode long strings of ones with shorter codes, thus transmit shorter codes (on average). For the third there is not much we can do, as it is random with equal probabilities of 0 and 1.

In all three instances, to compress the sequences it is much easier to construct a computation machine that carries out instructions one at a time, and instead of the sequence in question send the much shorter program.

The computation machine in question can be a Turing-complete computer, referred further down the text to as *universal computer.*

Definition A.75 (Universal Computer)**.** The function $\mathcal{U} : P \to S$ that maps a "program" $p \in P$ to its output $s \in S$ is called a universal computer.

Definition A.76 (Kolmogorov complexity)**.** Let x be a string from an alphabet $\aleph_X$. The Kolmogorov complexity of x with respect to a universal computer $\mathcal{U}$ is defined as

$$K_{\mathcal{U}}(x) = \min_{p:\mathcal{U}(p)=x} l(p)$$

If the length of x is known, then the conditional Kolmogorov complexity can be defined as

$$K_{\mathcal{U}}(x|l(x)) = \min_{p:\mathcal{U}(p,l(x))=x} l(p)$$

Theorem A.77 (Universality of Kolmogorov complexity)**.** *If $\mathcal{U}$ is a universal computer, for any other computer $\mathcal{A}$ there exists a constant $c_{\mathcal{A}}$ such that*

$$K_{\mathcal{U}}(x) \leq K_{\mathcal{A}}(x) + c_{\mathcal{A}}$$

for all strings x and the constant $c_{\mathcal{A}}$ does not depend on x.

Proof. For each computer $\mathcal{A}$ there is a program $s_{\mathcal{A}}$ that allows it to simulate any other computer $\mathcal{U}$. Its length $l(s_{\mathcal{A}}) = c_{\mathcal{A}}$, and

$$\mathcal{U}(s_{\mathcal{A}}p) = \mathcal{A}(p)$$

for all programs p, hence

$$\begin{aligned} K_{\mathcal{U}}(x) &= \min_{p:\mathcal{U}(p)=x} l(p) \\ &\leq \min_{p:\mathcal{U}(s_{\mathcal{A}}p)=x} l(s_{\mathcal{A}}p) \\ &= \min_{p:\mathcal{A}(p)=x} (l(p) + c_{\mathcal{A}}) \\ &= K_{\mathcal{A}}(x) + c_{\mathcal{A}}. \end{aligned}$$

The inequality comes from the fact that $\{s_{\mathcal{A}}p : p \in P\}$ is subset of all programs, and restricting the set may only increase the minimum. □

Because $c_{\mathcal{A}}$ depends only on the universal computer in question, and not on the sample, even for large $c_{\mathcal{A}}$ the effect will be negligible for long enough string x. That is why we can drop $\mathcal{U}$ and treat $\mathcal{K}(x)$ as if the specific computer is not important.

Note that this is problematic for finite strings x, where the complexity may be dominated by $c_{\mathcal{A}}$. Regrettably this is a serious problem for applications of Kolmogorov complexity in statistical inference, as any possible data is finite, and at times can be quite short.

Theorem A.78 (Lower bound of Kolmogorov complexity)**.** *The number of strings x with $K(x) < k$ satisfies*

$$\#\left\{x \in \{0,1\}^* : K(x) < k\right\} < 2^k$$

This theorem directly mirrors Kraft's inequality (Theorem A.53), and we can complete the equivalence by defining the so-called "universal probability" as the probability that a randomly chosen program will halt.

Lemma A.79. *For any computer $\mathcal{U}$, the* universal probability *of the string x is defined as*

$$P_{\mathcal{U}}(x) = \sum_{p:\mathcal{U}(p)=x} 2^{-l(p)} = \mathbb{P}(\mathcal{U}(P) = x).$$

This is the probability that a random program will produce the string x.

To complete the connection between between entropy and Kolmogorov complexity we can turn to the following theorem.

Theorem A.80 (Kolmogorov complexity vs entropy)**.** *Let the stochastic process $\{X_i\}$ be drawn according to the probability mass function $f(x)$ on the finite alphabet $\aleph$. Then there exists a constant c such as*

$$H(X) \leq \frac{1}{n} \sum_{x^n} f(x^n) K(x^n|n) \leq H(X) + \frac{(|\aleph| - 1) \log n}{n} + \frac{c}{n}$$

for all n.

Corollary A.81. *For an i.i.d. process* $\{X_i\}$,

$$\mathbb{E}\frac{1}{n}K(X^n|n) \to H(X).$$

This is a very strong connection between Kolmogorov complexity and entropy. In terms of mean code length it shows that Kolmogorov complexity cannot give us on average better codes than Shannon codes. That is a sufficient motivation to use Shannon codes, that is, to restrict ourselves to algorithms that correspond to probability distribution coding schemes, because they are so much simpler than universal computers, and asymptotically give us the same benefits.

In parallel to typical sets (A.35), algorithmically random sequences are defined as those whose complexity is close to their length.

Definition A.82 (Algorithmically random sequence)**.** A sequence $x_1, x_2, \ldots, x_n$ is said to be algorithmically random if

$$K(x_1 x_2 \ldots x_n|n) \geq n$$

This also means that those sequences cannot be compressed.

Definition A.83 (Incompressible sequences)**.** An infinite string x is called incompressible if

$$\lim_{n\to\infty} \frac{K(x_1 x_2 \ldots x_n|n)}{n} = 1$$

Theorem A.84. *If a string* $x_1 x_2 \ldots x_n$ *is incompressible, then it satisfies the law of large numbers so that*

$$\frac{1}{n}\sum_{i=1}^{n} x_i \to \frac{1}{2}$$

More generally, since any statistical test for uniformness can serve as a basis for a compression scheme, if it is not satisfied, incompressible sequences must satisfy them all.

Definition A.85 (Kolmogorov structure function)**.** The function $K_k(x^n|n)$ defined on binary strings $x \in \{0,1\}^n$ is defined as

$$K_k(x^n|n) = \min_{\substack{p\,:\,l(p)\leq k \\ \mathcal{U}(p,n)=S \\ x^n \in S \subseteq \{0,1\}^n}} \log|S|$$

The set S is the smallest set that can be described with no more than k bits and which includes x^n. $\mathcal{U}(p,n) = S$ signifies that running the program p with data n on the universal computer $\mathcal{U}$ will print out the indicator function of S.

Definition A.86 (Kolmogorov minimal sufficient statistic)**.** For a given small constant c, let k^* be the least k satisfying

$$K_k(x^n|n) + k \leq K(x^n|n) + c$$

Let S^{**} be the corresponding set and let p^{**} be the program that prints out the indicator function of S^{**}. Then we shall say that p^{**} is a *Kolmogorov minimal sufficient statistic* for x^n.

The connection between MDL and Kolmogorov minimal sufficient statistic can be found in Section 2.3.1, or see [Cover and Thomas, 2012], pp. 508.

B. Calculus of the Dirac δ-function

This appendix covers the basic properties of generalized functions, and in particular the δ-function. The exposition follows that of [Strichartz, 1994] and serves as introduction of the ideas that are required for rigorous definitions of the operations in Section 1.1.2, Chapter 3 and Chapter 4.

Theorems and definitions from [Strichartz, 1994] are referenced by number, whenever possible.

B.1. Definition and properties

Definition B.1 (Test function)**.** A function $\psi : \mathbb{R}^n \to \mathbb{R}$ is called a *test function* if it vanishes near the boundary of a compact set $M \subset \mathbb{R}^n$ and is zero outside it, while having continuous partial derivatives of every order.

The set of test functions is denoted by $\mathcal{D}$.

This allows us to define a generalized function, or distribution, by the effect it has on a test function:

Definition B.2 (Generalized function)**.** f is called a *generalized function*, if it belongs to the class of linear functionals operating on $\mathcal{D}$. The set of generalized functions is denoted by $\mathcal{D}'$.

The notation uses $\langle \cdot | \cdot \rangle : \mathcal{D}' \times \mathcal{D} \to \mathbb{R}$ to describe the functional's effect on a test function as a real value:

$$\langle f | \psi \rangle$$

and linear means that the following is true:

$$\langle f | \alpha\psi + \beta\phi \rangle = \alpha \langle f | \psi \rangle + \beta \langle f | \phi \rangle . \tag{B.1}$$

Other technical conditions are also required, such as continuity of the linear functional, but for the well-behaved functions that we will use we do not need to posit them.

Some generalized functions are in fact functions, and can be defined via Lebesgue integrals:

$$\langle f | \psi \rangle = \int f(x^n)\psi(x^n)dx^n = \int f(x^n)d\Psi(x^n). \tag{B.2}$$

In this case the linearity condition (B.1) is trivially satisfied.

A way to define generalized functions is to use continuity, using the following theorem.

Theorem B.3 ([Strichartz, 1994], Theorem 2.2.1, page 13)**.** *Given any generalized functions $f \in \mathcal{D}'$, there exists a sequence $\{\psi_k\}$ of test functions such that $\psi_k \to f$ as distribution, i.e. for any test function $\phi \in \mathcal{D}$ we have*

$$\langle \psi_k | \phi \rangle = \int \psi_k(x^n)\phi(x^n)dx^n \to \langle f|\phi\rangle \,. \tag{B.3}$$

On the basis of the theorem above we can define operations over distributions. We illustrate the way by defining the differential operator $\frac{d}{dx}$ for $n = 1$.

Corollary B.4. *Let $f \in \mathcal{D}'$. Then we can define $\frac{d}{dx}f$ as*

$$\Big\langle \frac{d}{dx} f|\phi\Big\rangle = -\Big\langle f|\frac{d}{dx}\phi\Big\rangle.$$

Proof. Define the sequence $\{\psi_k\}$ of test functions such that $\psi_k \to f$ as distribution using Theorem B.3. Then

$$\begin{aligned}
\Big\langle \frac{d}{dx} f|\phi\Big\rangle &= \lim_{k\to\infty} \int \frac{d}{dx}\left(\psi_k(x)\right)\phi(x)dx \\
&= \lim_{k\to\infty} \int \phi(x)d\psi_k(x) \\
&= \lim_{k\to\infty} \left[\phi(x)\psi_k(x)|_{-\infty}^{\infty} - \int \psi_k(x)d\phi(x)\right] \\
&= \lim_{k\to\infty} \left[-\int \psi_k(x)\left(\frac{d}{dx}\phi(x)\right)dx\right] \\
&= -\Big\langle f|\frac{d}{dx}\phi\Big\rangle
\end{aligned}$$

where the boundary terms are zero because ψ_k and ϕ are test functions. □

Thus to define operations over generalized functions we must consider how these operations affect the test function. This is the basis of the so-called adjoint identities, [Strichartz, 1994], Section 2.3:

Definition B.5. Let T and S be operators such that

$$\int T\psi(x^n)\phi(x^n)dx^n = \int \psi(x^n)S\phi(x^n)dx^n$$

for every $\psi, \phi \in \mathcal{D}$. Such T and S are called *adjoint identities*.

Trivially, if T has a corresponding adjoint identity S, using Theorem B.3 we can define Tf using

$$\Big\langle Tf|\phi\Big\rangle = \Big\langle f|S\phi\Big\rangle.$$

The converse is trivial.

We present the following basic properties without proof.

Lemma B.6 (Operations on distributions)**.** *Let $f \in \mathcal{D}'$. Then*

1. *If τ_r is translation of the argument, i.e. $\tau_r\psi(x^n) = \psi(x^n + r)$, then*
$$\Big\langle \tau_r f|\phi \Big\rangle = \Big\langle f|\tau_{-r}\phi \Big\rangle.$$
2. *If σ_α is scaling of the argument, i.e. $\sigma_\alpha\psi(x^n) = \psi(\alpha x^n)$ then*
$$\Big\langle \tau_\alpha f|\phi \Big\rangle = \Big\langle f|\tau_{\alpha^{-1}}\phi \Big\rangle.$$
3. *Scaling the function:*
$$\Big\langle \alpha f|\phi \Big\rangle = \alpha\Big\langle f|\phi \Big\rangle.$$
4. *The operator $\Big\langle \cdot|\cdot \Big\rangle$ is linear over its first argument as well, i.e. if $g \in \mathcal{D}'$ then*
$$\Big\langle f + g|\phi \Big\rangle = \Big\langle f|\phi \Big\rangle + \Big\langle g|\phi \Big\rangle$$

Using Theorem B.3, we will give two definitions of the Dirac δ-function.

Definition B.7 (Dirac δ-function)**.** The Dirac δ-function is defined as

$$\langle \delta|\phi \rangle = \phi(0). \tag{B.4}$$

Alternatively, define a sequence of test functions such that

1. $\psi_n \in \mathcal{D}^n$;
2. ψ_n has support on $\left[-\frac{1}{n}; \frac{1}{n}\right]^n$;
3. $\int \psi_n(x^n)dx^n = 1$.

Then $\psi_n \to \delta$ as distributions via Theorem B.3.

Alternative notation is to write (B.4) as

$$\int f(x^n)\delta(x^n)dx^n = f(0),$$

and this is not interpreted as the usual integral, but rather through (B.3).

Using lemma B.6, the δ-function can be proven to have the following properties:

Lemma B.8 (Properties of the delta function)**.**
1. $\delta(x^n + t)f(x^n) = f(x^n - t)$
2. $\delta(sx^n) = s^{-n}\delta(x^n)$*, so is homogeneous of degree* -1*.*
3. $\delta(g(x)) = \sum_{i=1}^n \frac{\delta(x-x_i)}{|\nabla g(x_i)|}$*, where x_i are all solutions of $g(x) = 0$.*

B.2. Probability distributions

Probability distributions are, unsurprisingly, generalized functions, and their effect on test functions is in fact the expectation of the function of the random variable.

Let f be the p.d.f. of a random variable X^n. Then

$$\mathbb{E}g(X^n) = \int g(x^n)f(x^n)dx^n = \langle f|g\rangle\,.$$

Let X be a discrete random variables taking values $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots$. The distribution of X can be then written as

$$f(x) = \sum_{i=1}^{\infty} p_i \delta(x - x_i). \tag{B.5}$$

In (B.5) f is not a function, but a generalized distribution, and we can still manipulate it like it is a density function.

The following theorem is very useful to represent the distribution of a function of random variables ([Kay, 1993], equation (5A.1)).

Theorem B.9 (Distribution of function)**.** *Let $g : \mathbb{R}^k \to \mathbb{R}$ be a function and X^k be a random variable with distribution $f_{X^n}(x^n)$. The distribution of $g(X^n)$ is denoted $f_{g(X^n)}$ and can be defined as*

$$f_{g(X^n)}(s) = f_{X^n}(x^n)\delta(g(x^n) - s). \tag{B.6}$$

One way to see the above is via properties of the δ-function, and the other is via a change of variables.

Proof. Using lemma B.8, point 3 we can see that looking for fixed $x_1, \ldots, x_{n-1}$ the distribution of $f_{X^n}(x^n)\delta(g(x^n) - s)$ takes the form

$$f_{X^n}(x^n)\left(\sum_{k=1}^{K} \frac{\delta(x - x_k)}{|\nabla \tilde{g}(x_k)|}\right) = \sum_{k=1}^{K} \frac{f_{X^n}(x_k^n)}{|\nabla \tilde{g}(x_k)|},$$

where $\tilde{g}(x^n) = g(x^n) - s$.

This however is exactly the form of the conditional distribution of $f_{g(X^n)}(s)$, see [Papoulis et al., 2002], pp 130, *Fundamental Theorem.* □

Additional benefit of using (B.6) to define conditional probability is that this way we explicitly define the measures, thus the conditional probability, if it exists, is regular, thus we can sidestep issues that may arise with reparameterization, e.g. Borel-Kolmogorov paradox.

In the main text of the dissertation, we will use the linearity of the δ-function in order to simplify the expressions and transformations. Before doing the actual numerical calculation, the δ-function can be removed via an appropriate change of variables.

C. Massively Parallel Particle Swarm Optimization

The aim and purpose of this section is to describe a relatively new method of numerical optimization, called Particle Swarm Optimization (PSO), its ideas and practical implementation.

To do the numerical computations in this dissertation, a custom implementation was created, called Massively Parallel Particle Swarm Optimization (MPPSO), available at `https://github.com/madcowbg/MPPSO`.

Section C.1 describes in broad strokes what the PSO strives to do and how it relates to similar concepts in computer science. Section C.2 describes the basic algorithm. Section C.3 describes the particular implementation used in the numerical calculation in Chapter 4.

C.1. Overview

Particle Swarm Optimization (PSO) is a novel method of numerical optimization introduced in [Kennedy and Eberhart, 1995]. The basic idea is that by using a "swarm" of particles with very simple behavior and letting them randomly traverse an n-dimensional space in search of what they consider the best location, we can solve highly nonlinear and even non-smooth optimization problems very quickly and reliably.

It can be viewed as a mid-level form of Artificial life (A-life) or biologically derived algorithm in that the construction was inspired by watching flocks of birds home-in on a target by sharing information between individuals, and is thus a group-order behavior rather than individual behavior that determines the outcome of the search.

PSO is obviously connected to evolutionary computations, as it lies inbetween genetic algorithms, where pieces of algorithms are combined and artificially selected from, and evolutionary programming, where the algorithms are fixed, but their parameters are left to evolve.

PSO is also dependent on stochastic processes for its innovations, much like evolutionary programming. The tendencies of individual birds to what they have seen as the best location and what they know that their flock mates consider as the best location is conceptually similar to the crossover operation utilized by genetic

algorithms. The mechanics of how the swarm actually evaluates the "fitness" of a particular individual will be described in Section C.2.

PSO as developed in [Kennedy and Eberhart, 1995] is a very simple concept, its practical implementation can be done in a few lines of computer code. The algorithm itself is uses only basic algebra, and is computationally inexpensive in terms of both memory requirements and speed. Early testing has found the implementation to be effective with several kinds of problems.

Furthermore, due to its nature of modeling the behavior of semi-independent individual particles, PSO is very highly parallelizable. With the recent advances of massively parallel computations on graphics processing units (GPUs), as well as the addition of simple and streamlined tools to program for them, interests in the application of PSO has been renewed.

Some notes on the details of the implementation of a PSO on a GPU, in particular when using Compute Unified Device Architecture (CUDA), can be found in [Iliana et al., 2011, Zhou et al., 2009].

Further explorations and refinements of the concepts are available for example in [Peer et al., 2003], as well as systematic exploration of the specific improvements in the PSO procedure, parameter tuning, etc in [Bergh, 2001].

The next section will describe the basic PSO algorithm in more details.

C.2. The Algorithm

The basis of the algorithm for PSO is the individual particle's behavior. Each particle has the following behaviour:

1. the current velocity of the particle;
2. a tendency towards its local maximum;
3. a tendency towards the swarm maximum.

The tendencies are weighted with random effects to get the acceleration and with this acceleration and a certain propensity to preserve the direction of movement (i.e. inertia) are averaged to determine the new velocity. This new velocity is used to determine the next position of the particle.

These influences can be seen on Figure C.1. The behavior can be summarized in the following formula from [Kennedy and Eberhart, 1995], where the variables v_i,x_i, y_i, etc. are n-dimensional vectors for ease of notation:

$$v_i(t+1) = wv_i(t) \tag{C.1}$$
$$+c_p r_{p,i}(t)(y_i - x_i(t)) \tag{C.2}$$
$$+c_g r_{g,i}(t)(\hat{y} - x_i(t)) \tag{C.3}$$

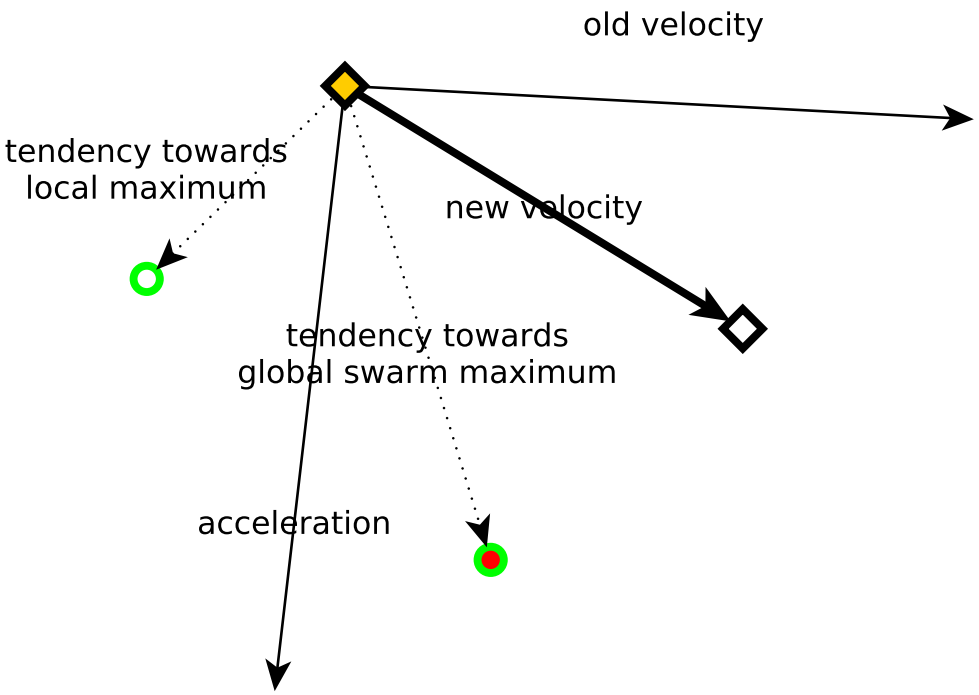

Figure C.1.: Basic individual particle behavior, with relevant components. Local (particle) maximum and global (swarm) maximum are defined as maxima so far and will change over time.

w is the inertia weight that determines how much the current direction of the particle is preserved. c_p and c_g are tuning constraints which show the strength of the tendencies towards the found so far personal best and the global best. y_i are the coordinates of the personal best, and $\hat{y}$ are the coordinates of the swarm best.

The randomness is encapsulated into $r_{p,i}(t)$ and $r_{p,i}(t)$ which are uniform and independent random variables with values in $[0, 1]$.

Another way to look at the above is by the design ideas that come from flock behaviour. In such terms:

- (C.1) is called the inertia of the individual;
- (C.2) is the cognitive component, representing nostalgia of the individual;
- (C.3) is the social component, representing envy of the other particles' success.

The algorithm can be summarized as follows:

Algorithm C.1 (PSO)**.** *For a given target function $f(x)$, the PSO algorithm is executed as follows:*

1. *A random population of P particles are initialized with random position $x_i(0)$, and random velocity $v_i(0)$ for all $i = 1, \ldots, P$.*
2. *Caclulation step: for all particles, $f(x_i(t))$ is calculated.*
3. *Reduction step: for all particles, their personal best y_i is updated, if needed. The global best $\hat{y}$ is updated, if needed.*
4. *Acceleration step: $r_{p,i}(t)$ and $r_{p,i}(t)$ are simulated, and $v_i(t+1)$ is calculated via (C.1)-(C.3).*
5. *Movement step:*

$$x_i(t+1) = x_i(t) + v_i(t+1).$$

6. *Until the criteria for convergence is met, or $t = T$ iterations have passed have been calculated, repeat steps 2-5.*

The tuning of the procedure above can be done in several ways, as is duscussed in [Peer et al., 2003] and [Bergh, 2001]. The following updates have been proposed by various authors with different aims in mind:

- Velocity is restricted to preven too fast convergence due to envy, i.e.

$$v_i^j(t) \leftarrow sign(v_i^j(t))max(|v_i^j(t)|, V_{max}^j).$$

- c_p and c_g can be tuned in order to speed up the convergence. In the seminal paper [Kennedy and Eberhart, 1995], the authors suggests using 2 for both (i.e. half the time the particle overshoots the target). Different values provide different trade-offs between nostalgia and envy.
- Introducing neighborhood topology - instead of using the global best, each particle uses the global best that it knows about, by restricting information flow between particles. Different topologies are possible, the most popular non-fully connected being ring topology where each particle has two neighbours.
- A more recent idea in [Maruyama and Saito, 2014] is to introduce collision mechanics to restrict the propensity of multiple particles to cluster at the same solution. Thus if multiple solutions are possible, e.g. if solutions are periodic, more than one of them can be found.

C.3. The Implementation

Based on the papers provided, a new implementation has been done for the calculations in this dissertation to more fully answer the requirements of the used algorithms. The code is packaged in a project called Massively Parallel Particle

Swarm Optimization (MPPSO) and is available for download in `https://github.com/madcowbg/MPPSO`.

This optimizer is used in the numerical calculations in Chapter 4 with a MATLAB integration via a compiled mex function.

The following PSO parameters can be changed:

- inertia - default is $w = 0.7$;
- c_p, c_g - defaults are $c_p = c_g = 1.4$;
- maximum velocity, default is $V_{max} = 2$;
- number of particles, default is $P = 128$;
- number of iterations, default is $T = 100$.

The only neighborhood topology currently supported is with total connectivity, i.e. every particle of the swarm is directly linked to every other particle.

It is advised that the number of particles is a large power of 2 due to the architecture of the GPU - CUDA uses Single Instruction Multiple Threads (SIMT) parallelism, which means that to obtain the benefits a large number of executions of the same instruction must be peformed by multiple threads.

The execution in current GPUs is performed in bursts of 32 threads, but with the advancing thread count it is advisable to use as many threads as possible, thus the defailt of $P = 128$.

The architecture class diagram is shown on Figure C.2. The top-level parallelisation is done with the class *ParallelPSO*, which is initialized with several PSO objects and spawns them in separate threads that use separate CUDA streams.

Each *PSO* object encapsulates one task to fit. *ParallelPSO* executes them in parallel, but since instructions are encapsulated into a *PSO* object, the *PSO* must have enough particles itself so that it can saturate the GPU in order to have the best performance.

The important requirement about PSO objects is that the bulk of their data is available on the GPU, denoted by gpuFloat arrays in Figure C.2. Small data pieces, like single double or floating point parameters can be transfered to the GPU during the kernel call.

GPU kernels are pieces of code that execute on the GPU. Custom GPU kernels are called by the PSO code to evaluate the fitness function of each particle (*Caclulation step*). Then Random Number Generator, Reduction routines and other generic GPU kernels are called by the inherited methods of the PSO to evolve the swarm (*Reduction step*, *Acceleration step*, *Movement step*).

Three example implementations of *PSO* are available in the code, all of which are concerned with the fitting and solving for x_n^*, x_{n-1}^* for the Student-T distribution. Their use in this dissertation is described in Chapter 4.

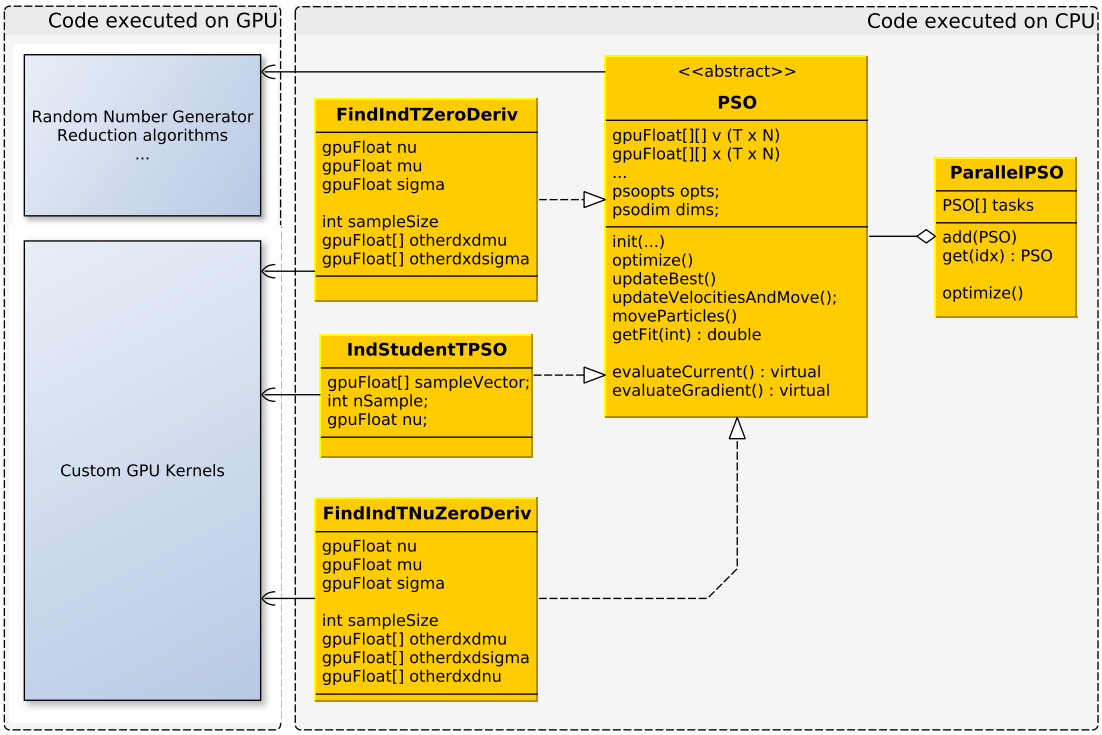

Figure C.2.: High-level architecture class diagram of MPPSO.

One additional prerequisite to get the most performance out of the GPU is that the GPU code should not branch much, e.g. via ifs or for cycles with different numbers of iterations. This is needed because the GPU handles branching by idling for threads that go on another branch, which can reduce the performance.

These and other ways to make sure that the GPU code is optimal for CUDA can be found in [Cook, 2013] and the CUDA documentation from NVIDIA.

Bibliography

[Adriaans and Vitanyi, 2007] Adriaans, P. and Vitanyi, P. (2007). The power and perils of MDL. In *International Symposium on Information Theory*, number 1, pages 2216–2220.

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6).

[Andrienko et al., 2000] Andrienko, Y. A., Brilliantov, N., and Kurths, J. (2000). Complexity of two-dimensional patterns. *The European Physical Journal B-Condensed Matter and Complex Systems*, 15(3):539–546.

[Asadi and Seyfe, 2013] Asadi, H. and Seyfe, B. (2013). Source Number Estimation via Entropy Estimation of Eigenvalues (EEE) in Gaussian and Non-Gaussian Noise. *arXiv preprint arXiv:1311.6051*, pages 1–18.

[Barron et al., 1998] Barron, A., Rissanen, J., and Yu, B. (1998). The Minimum Description Length Principle in Coding and Modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760.

[Bedrick and Tsai, 1994] Bedrick, E. and Tsai, C. (1994). Model selection for multivariate regression in small samples. *Biometrics*, 50(1):226–231.

[Bergh, 2001] Bergh, F. V. D. (2001). *An Analysis of Particle Swarm Optimizers*. PhD thesis.

[Blume, 1975] Blume, M. (1975). Betas and their regression tendencies. *The Journal of Finance*.

[Candès et al., 2011] Candès, E., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37.

[Cook, 2013] Cook, S. (2013). *CUDA Programming*. Elsevier.

[Corcoran et al., 2014] Corcoran, J., Tran, D., and Levine, N. (2014). An Efficient Search Strategy for Aggregation and Discretization of Attributes of Bayesian Networks Using Minimum Description Length. *arXiv preprint arXiv:1404.0752*.

[Cover and Thomas, 2012] Cover, T. and Thomas, J. (2012). *Elements of information theory*. Wiley.

[de Brauwere et al., 2005] de Brauwere, A., De Ridder, F., Pintelon, R., Elskens, M., Schoukens, J., and Baeyens, W. (2005). Model selection through a statistical analysis of the minimum of a weighted least squares cost function. *Chemometrics and Intelligent Laboratory Systems*, 76(2):163–173.

[Fishler et al., 2002] Fishler, E., Grosmann, M., and Messer, H. (2002). Detection of signals by information theoretic criteria: general asymptotic performance analysis. *IEEE Transactions on Signal Processing*, 50(5):1027–1036.

[Friedman and Goldszmidt, 1996] Friedman, N. and Goldszmidt, M. (1996). Discretizing continuous attributes while learning Bayesian networks. *International Conference on Machine Learning (ICML)*.

[Grunwald, 2004] Grunwald, P. (2004). A Tutorial Introduction to the Minimum Description Length Principle. *arXiv preprint math/0406077*.

[Grunwald, 2007] Grunwald, P. (2007). *The minimum description length principle*. The MIT Press.

[Grünwald and Kotlowski, 2010] Grünwald, P. and Kotlowski, W. (2010). Prequential plug-in codes that achieve optimal redundancy rates even if the model is wrong. *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 1383–1387.

[Grünwald et al., 2005] Grünwald, P., Myung, J. I., and Pitt, M. (2005). Advances in Minimum Description Length: Theory and Applications.

[Haddadi et al., 2010] Haddadi, F., Malek-Mohammadi, M., Nayebi, M. M., and Aref, M. R. (2010). Statistical Performance Analysis of MDL Source Enumeration in Array Processing. *IEEE Transactions on Signal Processing*, 58(1):452–457.

[Hamming, 1986] Hamming, R. W. (1986). *Coding and Information Theory*. Prentice Hall.

[Harremoës, 2013] Harremoës, P. (2013). Extendable MDL. *IEEE International Symposium on Information Theory Proceedings (ISIT)*, page 9.

[Hedayati and Bartlett, 2012a] Hedayati, F. and Bartlett, P. (2012a). The optimality of Jeffreys prior for online density estimation and the asymptotic normality of maximum likelihood estimators. *Proceedings of the Fifth Workshop on Information Theoretic Methods in Science and Engineering (WITMSE 2012)*, 23:1–13.

[Hedayati and Bartlett, 2012b] Hedayati, F. and Bartlett, P. L. (2012b). Exchangeability Characterizes Optimality of Sequential Normalized Maximum Likelihood and Bayesian Prediction with Jeffreys Prior. *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, 22:504–510.

[Iliana et al., 2011] Iliana, I., Liera, C., Antonio, M., and Castro, M. C. J. A. (2011). Parallel particle swarm optimization using GPGPU. *CIE*, pages 1–3.

[Kass and Raftery, 1995] Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

[Kay, 1993] Kay, S. (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory, 1993*, volume 1.

[Kennedy and Eberhart, 1995] Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In *IEEE International Conference on Neural Networks*, volume 4, pages 1942–1948. Ieee.

[Kolmogorov, 1963] Kolmogorov, A. N. (1963). On Tables of Random Numbers. *Sankhya: The Indian Journal of Statistics, Series A*, 25(4):369–376.

[Konishi and Kitagawa, 2008] Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer Science & Business Media.

[Kotlowski and Grunwald, 2011] Kotlowski, W. and Grunwald, P. (2011). Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. *Conference on Learning Theory (COLT)*, 19:457–475.

[Kotz and Nadarajah, 2004] Kotz, S. and Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge University Press, Cambridge, New York, Madrid.

[Li and Vitanyi, 1993] Li, M. and Vitanyi, P. (1993). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag.

[Liski and Liski, 2008a] Liski, A. and Liski, E. (2008a). MDL knot selection for penalized splines. *Proceedings of the First Workshop on Information Theoretic Methods in Science and Engineering*, (2):3–4.

[Liski and Liski, 2008b] Liski, E. and Liski, A. (2008b). Model selection in linear mixed models using MDL criterion with an application to spline smoothing. *Proceedings of the First Workshop on Information Theoretic Methods in Science and Engineering (WITMSE 2008)*, 3:1–2.

[Loh et al., 1992] Loh, W.-Y., Fang, K.-T., Kotz, S., and Ng, K.-W. (1992). Symmetric Multivariate and Related Distributions. *Technometrics*, 34:235.

[Malyutov et al., 2013a] Malyutov, M., Li, X., Zhang, T., and Li, Y. (2013a). Statistics via VLMC-modeling: Three case studies.

[Malyutov et al., 2013b] Malyutov, M., Zhang, T., Li, Y., and Li, X. (2013b). Time series homogeneity tests via VLMC training. *Information Processes*, 13(4):401–441.

[Maruyama and Saito, 2014] Maruyama, K. and Saito, T. (2014). A Collision PSO for Search of Periodic Points. In *Nonlinear Dynamics of Electronic Systems*, pages 59–67.

[McQuarrie et al., 1997] McQuarrie, A., Shumway, R., and Tsai, C. (1997). The model selection criterion AICu. *Statistics & probability letters*, 34(3):285–292.

[Nadler, 2010] Nadler, B. (2010). Nonparametric Detection of Signals by Information Theoretic Criteria: Performance Analysis and an Improved Estimator. *IEEE Transactions on Signal Processing*, 58(5):2746–2756.

[Nonchev, 2013a] Nonchev, B. (2013a). Minimum description length principle and distribution complexity of spherical distributions. *18th European Young Statisticians Meeting*, pages 33–37.

[Nonchev, 2013b] Nonchev, B. (2013b). Minimum Description Length Principle in Discriminating Marginal Distributions. *Pliska Studia Mathematica Bulgarica*, 22(125):101–114.

[Nonchev, 2014] Nonchev, B. (2014). Minimum Description Length Principle for Fat-Tailed Distributions. In Mladenov, V. M. and Ivanov, P. C., editors, *Nonlinear Dynamics of Electronic Systems*, volume 438 of *Communications in Computer and Information Science*, pages 68–75. Springer International Publishing, Cham.

[Pandey and Dukkipati, 2013] Pandey, G. and Dukkipati, A. (2013). Minimum description length principle for maximum entropy model selection. *Information Theory Proceedings (ISIT)*.

[Papoulis et al., 2002] Papoulis, A., Pillai, U., and Pillai, S. U. (2002). *Probability, random variables and stochastic processes*.

[Peer et al., 2003] Peer, E., van den Bergh, F., and Engelbrecht, A. (2003). Using neighbourhoods with the guaranteed convergence PSO. In *Proceedings of the 2003 IEEE Swarm Intelligence Symposium. SIS'03*.

[Rachev et al., 2005] Rachev, S. T., Menn, C., and Fabozzi, F. J. (2005). *Fat-Tailed and Skewed Asset Return Distributions*. John Wiley & Sons.

[Ramírez and Sapiro, 2012a] Ramírez, I. and Sapiro, G. (2012a). An MDL framework for sparse coding and dictionary learning. *IEEE Transactions on Signal Processing*, 60(6):2913–2927.

[Ramírez and Sapiro, 2012b] Ramírez, I. and Sapiro, G. (2012b). Low-rank data modeling via the Minimum Description Length principle. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2165–2168.

[Ridder and Pintelon, 2005] Ridder, F. D. and Pintelon, R. (2005). Modified AIC and MDL model selection criteria for short data records. *IEEE Transactions on Instrumentation and Measurement*, (1):1–19.

[Rissanen, 1978] Rissanen, J. (1978). Modeling by shortest data description.

[Rissanen, 1983] Rissanen, J. (1983). A universal data compression system. *IEEE Transactions on Information Theory*, 29(5):656–664.

[Rissanen, 1999] Rissanen, J. (1999). Hypothesis Selection and Testing by the MDL Principle. *The Computer Journal*, 42(4):260–269.

[Rissanen, 2000] Rissanen, J. (2000). MDL Denoising. *IEEE Transactions on Information Theory*, 46(7):2537–2543.

[Rissanen, 2007] Rissanen, J. (2007). *Information and Complexity in Statistical Modeling (Information Science and Statistics)*. Springer.

[Rissanen and Roos, 2007] Rissanen, J. and Roos, T. (2007). Conditional NML universal models. *Information Theory and Applications Workshop*, (4):337–341.

[Rissannen, 1989] Rissannen, J. (1989). *Stochastic Complexity in Statisctical inquiry*. World Scientific.

[Roos and Rissanen, 2008] Roos, T. and Rissanen, J. (2008). On sequentially normalized maximum likelihood models. *Workshop on Information Theoretic Methods in Science and Engineering*, 68(1).

[Ryabko et al., 2010] Ryabko, B., Astola, J., and Malyutov, M. (2010). Compression-Based Methods of Prediction and Statistical Analysis of Time Series: Theory and Applications. *Tampere, TICSP series*, (56).

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[Shtarkov, 1987] Shtarkov, Y. (1987). Universal Sequential Coding of Single Messages. *Problems of Information Transmission*, 23(3):175–186.

[Stine and Foster, 2001] Stine, R. and Foster, D. (2001). The Competitive Complexity Ratio. In *Proceedings of the 2001 Conference on Information Sciences and Systems. WP8 1-6*.

[Strichartz, 1994] Strichartz, R. S. (1994). *A Guide to Distribution Theory and Fourier Transforms*. CRC Press.

[Viola, 2010] Viola, E. (2010). The complexity of distributions. In *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 202–211.

[Vitányi and Li, 2000] Vitányi, P. and Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *Information Theory, IEEE Transactions on*, 46(2):446–464.

[Zellner, 1976] Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms. *Journal of the American Statistical Association*, 71(354).

[Zenil and Soler-Toscano, 2012] Zenil, H. and Soler-Toscano, F. (2012). Two-dimensional Kolmogorov complexity and validation of the coding theorem method by compressibility. *arXiv preprint arXiv:1212.6745*, page 37.

[Zhou et al., 2009] Zhou, Y., Tan, Y., and Member, S. (2009). GPU-based Parallel Particle Swarm Optimization. (2):1493–1500.

Index