MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
ÚSTAV TEORETICKÉ FYZIKY A ASTROFYZIKY

Bakalářská práce

BRNO 2026 BC. SÁRA HASÍKOVÁ

MASARYKOVA
UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
ÚSTAV TEORETICKÉ FYZIKY A ASTROFYZIKY

Statistické projevy počáteční hmotnostní funkce hvězd od hvězdokup po částečně rozlišené populace

Bakalářská práce

Bc. Sára Hasíková

Vedoucí práce: Dr. rer. nat. Tereza Jeřábková **Brno 2026**

Bibliografický záznam

Autor:	Bc. Sára Hasíková Přírodovědecká fakulta, Masarykova univerzita Ústav teoretické fyziky a astrofyziky
Název práce:	Statistické projevy počáteční hmotnostní funkce hvězd od hvězdokup po částečně rozlišené populace
Studijní program:	Fyzika
Studijní obor:	Astrofyzika
Vedoucí práce:	Dr. rer. nat. Tereza Jeřábková
Akademický rok:	2025/2026
Počet stran:	xv + 60
Klíčová slova:	Počáteční hmotnostní funkce hvězd; IMF; hvězdokupy; hvězdné populace; random sampling; optimal sampling; syntetické hvězdokupy; machine learning; Fano faktor; částečně rozlišené populace;

Bibliographic Entry

Author:	Bc. Sára Hasíková Faculty of Science, Masaryk University Department of Theoretical Physics and Astrophysics
Title of Thesis:	Statistical Signatures of Stellar Initial Mass Function Sampling: From Star Clusters to Semi-Resolved Populations
Degree Programme:	Physics
Field of Study:	Astrophysics
Supervisor:	Dr. rer. nat. Tereza Jeřábková
Academic Year:	2025/2026
Number of Pages:	xv + 60
Keywords:	Initial mass function; IMF; star clusters; stellar populations; random sampling; optimal sampling; synthetic star clusters; machine learning; Fano factor; semi-resolved populations

Abstrakt

V této bakalářské práci se věnujeme samplingu počáteční hmotnostní funkce, tedy IMF, která tvoří jeden ze základních kamenů astrofyziky popisující rozložení hmotností hvězd při jejich zrodu. Blíže se zabýváme dvěma způsoby její realizace: random samplingem a optimal samplingem, které lze chápat jako dva vzájemné protipóly. Random sampling představuje stochastický přístup, při němž jsou hvězdné hmotnosti vybírány z IMF náhodně, což může vést k výraznému rozptylu mezi jednotlivými realizacemi. Optimal sampling naproti tomu chápeme jako zcela deterministický přístup bez Poissonova šumu. Cílem této práce je porovnat tyto dva přístupy, prozkoumat blíže jejich statistické vlastnosti a zjistit, zda je možné je od sebe rozeznat na základě vlastností výsledných hvězdných populací. K analýze využíváme syntetická data a metody strojového učení.

Abstract

In this thesis, we study the sampling of the initial mass function, or IMF, which is one of the cornerstones of astrophysics, describing the distribution of stellar masses at birth. We focus on two methods of its realization: random sampling and optimal sampling, which can be understood as two opposites. Random sampling represents a stochastic approach in which stellar masses are drawn randomly from the IMF, which can lead to significant scatter between individual realizations. Optimal sampling, on the other hand, is understood as a fully deterministic approach without Poisson noise. The aim of this thesis is to compare these two approaches, examine their statistical properties in more detail, and determine whether it is possible to distinguish them from each other based on the properties of the resulting stellar populations. We use synthetic data and machine learning methods for the analysis.

MUNI
SCI

MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
KOTLÁŘSKÁ 2, 611 37 BRNO
IČ: 00216224
DIČ: CZ00216224

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Akademický rok: 2025/2026

Ústav:	Ústav teoretické fyziky a astrofyziky
Studentka:	Bc. Sára Hasíková
Program:	Fyzika
Specializace:	Astrofyzika

Ředitel *ústavu* PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje bakalářskou práci s názvem:

Název práce:	Statistické projevy počáteční hmotnostní funkce hvězd od hvězdokup po částečně rozlišené populace
Název práce anglicky:	Statistical Signatures of Stellar Initial Mass Function Sampling: From Star Clusters to Semi-Resolved Populations
Jazyk práce:	angličtina

Oficiální zadání:
The stellar initial mass function (IMF) is a cornerstone of astrophysics, describing the distribution of stellar masses at birth. While its functional form is often assumed to be universal, the way in which stars populate it in individual systems may vary according to the physical conditions of star formation and the sampling process itself. This project will explore the statistical properties of different IMF sampling approaches, focusing on purely random sampling and ”optimal” sampling schemes, and investigate observational diagnostics capable of distinguishing between them. The student may work with synthetic star clusters generated via Monte Carlo methods and/or N-body simulations, analyse statistical trends in stellar content. Time permitting the developed methods can be tested against real data from the Gaia mission. Depending on interest, the project can be extended to the regime of semi-resolved stellar populations - where individual bright stars are resolved but the bulk of the light is unresolved - and compared to predictions from recent literature. The aim is to link statistical predictions to observable quantities that can be measured in star clusters or galaxies.

The IMF determines the number of low-mass and high-mass stars formed in a stellar population, directly influencing its luminosity, chemical enrichment, and dynamical evolution. Two widely discussed sampling paradigms are:
1. Random sampling, where stellar masses are drawn stochastically from the IMF;
2. Optimal sampling, where the most massive star and the entire mass spectrum are set by deterministic relations with the total stellar mass formed.

This project will begin with a review and statistical comparison of these methods, using simulated clusters to measure differences in mass distributions, luminosity functions, and other possible indicators. The work will include:
- Implementing or adapting code to generate synthetic clusters under different sampling schemes.
- Quantifying statistical scatter in key observables and exploring how it scales with cluster mass.
- Assessing whether quantities beyond the mass of the most massive star - e.g., number of high-mass stars, integrated colours, ionizing flux - can serve as diagnostics.
- (Optional) Comparing predictions to Gaia-based observations of young clusters.
- (Optional) Extending the analysis to semi-resolved populations (e.g. nearby galaxies), building on recent work by Ignacio Martín-Navarro and collaborators.
- (Optional) Incorporating N-body simulations to examine dynamical effects.

The exact direction will be refined according to the student's interests and the available data/tools. The project is suited to a mathematically minded student and offers a mix of statistical analysis, programming, and astrophysical interpretation.

Vedoucí práce:	Dr. rer. nat. Tereza Jeřábková
Datum zadání práce:	10. 9. 2025
V Brně dne:	5. 5. 2026

Zadání bylo schváleno prostřednictvím IS MU.

Bc. Sára Hasíková, 11. 9. 2025
Dr. rer. nat. Tereza Jeřábková, 24. 11. 2025
RNDr. Luboš Poláček, 11. 12. 2025

Poděkování

Na tomto místě bych chtěla poděkovat všem, kteří mi pomohli k úspěšnému dokončení této práce. Můj hlavní dík patří mé vedoucí Dr. rer. nat. Tereze Jeřábkové, která si i přes svůj nabitý harmonogram na mě vždy našla potřebný čas. Děkuji jí za odborné vedení, přátelský přístup, otevřenost, s jakou mě do celého tématu zasvětila, a za možnosti, které mi při psaní této práce poskytla. Dále bych chtěla poděkovat své rodině, kamarádům a všem blízkým, kteří mě vždy podporovali a pochopili, když jsem potřebovala čas na psaní. Děkuji všem, kteří věří, že mířím ke hvězdám.

Prohlášení

Prohlašuji, že jsem svoji bakalářskou práci vypracovala samostatně pod vedením vedoucího práce s využitím informačních zdrojů, které jsou v práci citovány. Zároveň prohlašuji, že jsem využila nástroje ChatGPT a Gemini pro jazykovou úpravu textu, konzultaci formulací a pomoc při kontrole a ladění kódu. Po použití těchto nástrojů autorka provedla kontrolu obsahu a přebírá za něj plnou zodpovědnost.

Brno 12. května 2026 .....

Bc. Sára Hasíková

Contents

List of Symbols

$\xi(m)$	Initial mass function
$\xi_L(m)$	Logarithmic form of the IMF
m	Stellar mass
$M_\odot$	Solar mass
$M_{\rm cl}$	Total cluster mass
$m_{\rm max}$	Maximum stellar mass
$m^*_{\rm max}$	Physical upper stellar mass limit
m_L	Lower stellar mass limit
m_i	Mass of the i-th star
dN	Number of stars in a mass interval
N	Total number of stars
$N_{{\rm bin},i}$	Number of stars in the i-th mass bin
$N_{\rm tot}$	Total number of stars used for normalization
f_i	Number fraction in the i-th mass bin
α	IMF power-law exponent
x	Salpeter exponent
k_i	Normalization constant of the i-th IMF segment
$\psi(t)$	Star formation rate
$\Psi(M_V)$	Luminosity function
$\tau(m)$	Stellar lifetime
M_V	Absolute visual magnitude
L	Luminosity
$L_{\rm max}$	Luminosity of the brightest star
$L_{\rm tot}$	Total luminosity
F	Fano factor
μ	Mean value
σ^2	Variance
D	Kolmogorov–Smirnov distance
ρ	Spearman correlation coefficient
p	p-value

Introduction

Stars have adorned our sky night after night since time immemorial. The sky is dotted with them like freckles on a face, like diamonds on a woman's neck, like drops of morning dew on grass. And although we could write poems and compose odes about stars, if we set aside all metaphors and poetic language, from a physical point of view, they are self-gravitating spheres of hot plasma that produce their own light and heat (of course, this does not detract from their beauty and charm). They are a constant object of interest for astrophysicists, as understanding their structure and the physical phenomena occurring within them opens the door to understanding the universe as a whole.

One of the most important parameters of a star is its mass. It is not just a number; mass plays a fundamental role in stellar physics, determining the evolution of star clusters and galaxies, influencing a star's luminosity, lifetime, and final fate. (Let us not give it all the credit, however; stellar evolution is also influenced by metallicity, rotation, binarity, the influence of the surrounding environment, and other factors.)

In this work, we focus on the Initial Mass Function (IMF), which describes the initial mass distribution of stars within a stellar population, a distribution that is reflected not only in the evolution of star clusters themselves but also in entire galaxies. When describing entire galaxies, however, it is not sufficient to consider only an isolated star cluster. A galaxy consists of many star clusters spanning a broad range of masses, and the resulting mass distribution of stars therefore depends not only on the IMF itself, but also on the mass distribution of the star clusters and the star formation rate. This problem is addressed by the concept of the Integrated Galactic Initial Mass Function (IGIMF), which extends the local IMF on galactic scales. The IGIMF thus represents the link between the stellar mass distribution of stars in star clusters and the resulting distribution of stars throughout the galaxy.

Since the IMF is a very broad topic, we have decided to focus only on a part of this issue, namely the realization of the IMF at the star cluster level. In other words, we are not concerned solely with the shape of the IMF as a continuous function, but primarily with the question of how this function gives rise to a specific set of stars with given masses. There is no single answer to this question, and as our knowledge and computational capabilities have grown, more methods for selecting stars from the IMF have emerged. This selection can significantly influence the properties of the resulting star cluster, especially in the case of star clusters with low total mass, since the presence or absence of a single massive star can alter the total number of stars, the cluster's luminosity, and its subsequent impact on the surrounding environment.

In this thesis, we compare two distinct approaches to implementing the IMF, which can be viewed as two extreme cases: random sampling and optimal sampling. Random

sampling treats the IMF as a probability distribution and selects stellar masses randomly from that distribution, which naturally leads to fluctuations between individual realizations. In practice, however, we often fix the total cluster mass $M_{\rm cl}$. As a result, the population we obtain is a constrained stochastic realization rather than a set of completely independent random draws. Optimal sampling, on the other hand, represents a deterministic approach in which a single specific distribution of stellar masses is determined for a given total mass of the star cluster. This approach suppresses random fluctuations and attempts to reproduce the theoretical IMF distribution as accurately as possible. Consequently, the mass of the most massive star, $m_{\rm max}$, is directly linked to the total cluster mass $M_{\rm cl}$, yielding a deterministic $m_{\rm max}$–$M_{\rm cl}$ relation.

We are primarily interested in these approaches from a statistical perspective. Throughout the study, we work with numerically generated datasets produced using our own (or substantially modified) functions to both generate and analyze model star clusters. We then systematically compare the two approaches for star clusters spanning a range of total masses and investigate how their differences manifest, for example, in the maximum stellar mass, the total number of stars, and the overall shape of the stellar mass distribution. For the reader, we provide a link to the publicly available Python scripts, datasets, and Jupyter notebooks stored on GitHub: https://github.com/Sarrisis/imf-sampling.git. Here, we view both random and optimal sampling primarily as statistical procedures for implementing the IMF rather than as complete physical models of star formation. We do not incorporate the full physics of gas fragmentation, accretion, feedback, magnetic fields, or dynamical evolution, factors that must be taken into account in applications to real observations.

The thesis is divided into four chapters. In Chapter 1, we introduce the reader to the theory of the IMF, outline its historical background, and describe the gradual development of this field. We also present several key questions that astrophysicists are still seeking to answer, such as the universality of the IMF, its physical origin, and how it manifests in specific stellar populations.

In Chapter 2, we turn to the first of the two approaches under comparison: random sampling. In this framework, we interpret the IMF as a probability distribution from which stellar masses are drawn at random. This perspective relies on several assumptions, such as the independence of individual draws or the absence of global constraints. We show that when a fixed total mass $M_{\rm cl}$ is introduced, these assumptions are partially violated, and we examine the resulting statistical consequences.

In Chapter 3, we introduce the optimal sampling approach, describe its mechanism, and analyze the features that distinguish it from random sampling. We focus in particular on the deterministic nature of this approach, the relationship between $m_{\rm max}$ and $M_{\rm cl}$, and how the absence of stochastic fluctuations affects the resulting properties of the star cluster.

In the final Chapter 4, we address the question of whether it is possible to detect the "signature" of a specific sampling method in simulated data. First, we analyze the statistical properties of the generated datasets, such as deviations from Poissonian behavior and correlations between mass bins, thereby building on the implications of imposing a fixed total mass $M_{\rm cl}$ discussed in Chapter 2. At the end of the chapter, we then use machine learning methods to test to what extent the underlying sampling method can be inferred retrospectively from selected properties of the stellar population.

Chapter 1

The IMF... What We Know and Do Not Know

The initial mass function (IMF) is an empirical function describing the initial mass distribution for a population of stars at the moment of their formation. The IMF links processes occurring on small scales in molecular clouds with the evolution of galaxies and the cosmological cycle of matter. It is key to interpreting stellar populations, their luminosity, chemical enrichment, and the formation of compact objects.

Mathematically, the function can be expressed simply as:

$$\mathrm{d}N = \xi(m)\,\mathrm{d}m, \tag{1.1}$$

where $\xi(m)$ denotes the number of stars formed per unit mass interval. Later in this chapter, we will show its more detailed derivation, although the mathematical form appears simple, its physical origin and implications are complex. Despite decades of observational and theoretical studies, the physical origin of the IMF is still not fully understood.

The IMF is relevant to many areas of astrophysics. As discussed by Kroupa [1], the IMF plays a crucial role in determining:

- the luminosity of the stellar system
- its mass locked up in long-lived late-type stars
- rate of core-collapse supernovae per late-type star
- rate of merging white dwarf binaries (SNIa) per late-type star
- rate of merging neutron star binaries per late-type star
- rate of merging black holes per late-type star
- energy and chemical element input into the interstellar medium

Over the past few decades, numerous studies have been conducted examining the IMF in terms of its shape, physical origin, and various properties. As illustrated in Fig. 1.1, a variety of functional forms have been proposed. As time passed, different astrophysicists have introduced alternative parameterizations. In this work, we adopt the Kroupa (2001) IMF [2]. However, we will discuss the general description of the other IMF forms in the next subsection 1.1 in order to summarize its development and history.

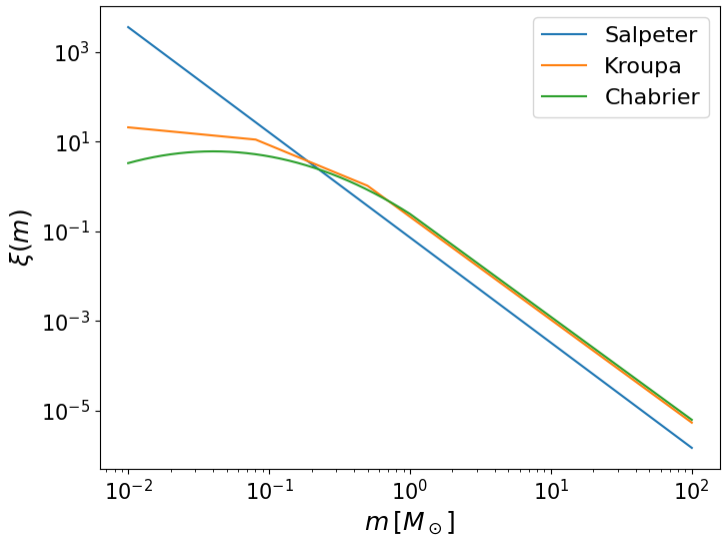


Figure 1.1: Salpeter, Kroupa, and Chabrier IMFs. The function $\xi(m)$ is plotted as a function of stellar mass m in the range 0.01–100$M_\odot$. For comparison, all curves are normalized over the same mass interval by requiring $\int m\xi(m)\,dm = 1$. The differences between the curves primarily reflect their distinct behavior in the low-mass range and the presence or absence of breaks or a log-normal transition. This figure is illustrative only; in the following chapters, we adopt a different mass range and normalization for the simulations.

1.1 A Little Bit of History

1.1.1 Salpeter IMF

The first mention of the IMF, or rather the introduction of the function and its initial form, was made by Edwin E. Salpeter in his 1955 work [3]. His life and broader academic career are described in more detail in [4, 5]. Here, we provide only a brief summary. Originally, he specialized in other areas of physics, particularly quantum electrodynamics and nuclear physics (perhaps the Bethe–Salpeter equation[1] sounds familiar?). Salpeter later turned his attention to astrophysics, where he studied the distribution of stellar luminosities and their evolution, which led to his seminal work *The Luminosity Function and Stellar Evolution*. This work then represented the first imaginary bridge between quantum mechanics and cosmology.

Salpeter sought to determine how many massive stars were born in our Galaxy and how many had already died since its formation. He wanted to derive the star birthrate function, i.e., the probability with which a star of a certain mass will be born at a certain time. His work was based on the statistics available at the time on the number of stars in the galactic disk, because these stars constituted the best-documented part of the visible population.

To derive the IMF, he used the overall observed distribution of stars by luminosity,

[1] The Bethe-Salpeter equation was published in 1951 by Hans Bethe and Edwin E. Salpeter. It describes the bound states of two relativistic particles in quantum theory.

expressed as the number of stars. He corrected this distribution for the influence of stars that had already left the main sequence and for the contribution of white dwarfs, thereby obtaining the so-called original luminosity function. He then used the relationship between luminosity and stellar mass to convert the luminosity distribution into a mass distribution and defined

$$\mathrm{d}N = \xi(m)\mathrm{d}(\log_{10} m) \tag{1.2}$$

as the number of stars in a logarithmic mass interval. The resulting equation was

$$\xi(m) \approx 0.03(m/M_{\odot})^{-1.35} \tag{1.3}$$

which we now know as the Salpeter IMF with power-law index x = 1.35, valid for $0.4 < m/M_{\odot} < 10$. Salpeter assumed that stars form at a constant rate, do not change their mass during their evolution, and that we know their lifetimes. He also assumed that the IMF remains constant over time. The graphical representation of the Salpeter IMF is shown in Figure 1.1 as a simple blue line. We can see that, unlike later parameterizations, it does not contain breaks.

He worked with the logarithmic form, but in the literature, it is also expressed in linear mass units with an index $\alpha = 2.35$, that is

$$\frac{\mathrm{d}N}{\mathrm{d}m} \propto m^{-2.35} \tag{1.4}$$

for stars with $m \gtrsim 1\,M_{\odot}$.

1.1.2 Miller and Scalo IMF

Later work showed that the IMF has a more complex shape, especially in the low-mass range, where changes in the internal structure of stars, opacity, and dominant physical processes are evident. This led to the introduction of multi-piece power-law or lognormal parameterizations of the IMF.

A more complex form of the IMF was first proposed by Miller and Scalo in 1979. The basis of their work [6] is the present-day mass function (PDMF) of main-sequence field stars in the solar neighborhood, which is defined as the number of main-sequence stars per unit logarithmic mass interval per square parsec in the solar neighborhood. In simple terms, the PDMF is a function that describes the current distribution of stars by mass.

The PDMF of main-sequence field stars $\phi_{\mathrm{ms}}(\log M)$ is related to the luminosity function of field stars $\phi(M_V)$, which describes the number of stars per unit absolute magnitude and volume in the Galactic disk. In their article, the authors emphasize that the IMF cannot be measured directly; instead, its determination depends on the PDMF, and this PDMF is derived from the luminosity function and corrected for observational and evolutionary effects.

They point out that the PDMF may be influenced by a number of factors and uncertainties, which they have taken into account in their analysis. The determination of the present-day mass function is also limited to a finite mass range, typically from about 0.1 to several tens of solar masses, with increasing uncertainties toward both low and high masses due to observational limitations.

However, the PDMF alone is not sufficient to determine the IMF, since it describes stars only up to the point of their evolution and thus contains more low-mass stars and fewer high-mass stars than there were originally. Based on this, Miller and Scalo established a relationship between the PDMF, the IMF, and the birthrate. The number of stars observed today depends not only on the initial mass distribution, but also on the time of their formation and their lifetimes. In a simplified notation, and assuming a consistent definition of the mass interval for both the PDMF and the IMF, this relation can be written schematically as

$$\mathrm{PDMF}(m) = \int_{t_0-\tau(m)}^{t_0} \xi(m)\,\psi(t)\,\mathrm{d}t, \tag{1.5}$$

where $\xi(m)$ is the initial mass function, $\psi(t)$ is the star formation rate, and $\tau(m)$ is the lifetime of a star of mass m. This notation represents a simplified form of the relations presented in [6] and is included here primarily to provide a general overview. In this expression (1.5), the PDMF denotes a general form of the present-day mass function; however, in practice, it is constrained using observations of main-sequence stars.

They tested various forms of the birthrate and concluded that only certain combinations, together with the PDMF, yield a meaningful IMF. Thus, the IMF must be smooth and free of unrealistic jumps. This creates a mutual constraint between the birthrate and the IMF. The IMF depends on the birthrate, but at the same time, the birthrate can be controlled through the IMF requirements.

Their results demonstrate that the IMF cannot be described by a single power-law, as originally proposed by Salpeter. While a power-law approximation remains valid at higher masses, the IMF flattens toward lower masses and exhibits a turnover, indicating a more complex, curved shape that can be approximated by a half-Gaussian distribution in $\log M$ over the mass range from $0.1\,M_\odot$ to about $50\,M_\odot$. In doing so, Miller and Scalo, like Salpeter, assumed that the IMF is time-independent and continuous in mass.

1.1.3 Kroupa IMF

Another significant extension and refinement of the IMF, which is frequently discussed and widely used, is presented by Kroupa in [2]. Kroupa's work builds on earlier studies by Scalo, which showed that the IMF cannot be described by a single power-law. By accounting for observational biases, particularly unresolved multiple systems, Kroupa derived a simpler and more consistent multi-part power-law form of the IMF. Although [2] is not the work in which this form of the multi-part power-law IMF was first introduced, it is the work in which its form is summarized and its variability is examined. This form of the IMF is also important for this thesis since (as we mentioned in the introduction to this chapter) we use it in the practical section for modeling and sampling star clusters.

The IMF is commonly inferred from observations by relating the luminosity function to stellar masses through a mass–luminosity relation (see, for example, Kroupa [2] and Kroupa et al. [7]). The luminosity function is defined through

$$\mathrm{d}N = \Psi\,\mathrm{d}M_V \tag{1.6}$$

which gives the number of stars in the magnitude interval $M_V \in [M_V, M_V + \mathrm{d}M_V]$. This represents the observable quantity. However, we would like to be able to calculate it using

the IMF, that is

$$\mathrm{d}N = \xi(m)\,\mathrm{d}m \tag{1.7}$$

as the number of stars with initial mass $m \in [m, m+\mathrm{d}m]$. These relationships (1.6) and (1.7) are, of course, related; we can write them as

$$\frac{\mathrm{d}N}{\mathrm{d}M_V} = -\frac{\mathrm{d}m}{\mathrm{d}M_V}\frac{\mathrm{d}N}{\mathrm{d}m}, \tag{1.8}$$

which leads to the equation

$$\Psi(M_V) = -\frac{\mathrm{d}m}{\mathrm{d}M_V}\xi(m). \tag{1.9}$$

It is the $\mathrm{d}m/\mathrm{d}M_V$ term that indicates the slope. It is the derivative of the stellar mass-luminosity relation and the important (and a bit problematic) part.

Kroupa's modification of the multi-part power-law form of the IMF arose as a result of correcting systematic observational biases. As we know, a large proportion of stars are found in binary or multiple systems. However, these systems often appear as single stars in observations, leading to an underestimation of the number of low-mass stars in particular, and thus to a distortion of the resulting IMF. After accounting for these effects, it turns out that the IMF cannot be described by a single power-law function, but rather exhibits changes in slope. Based on these corrections, Kroupa introduced an IMF with two changes in slope at masses $0.08\,M_\odot$ and $0.5\,M_\odot$.

If we cite the IMF as formulated in [2], its form is

$$\xi(m) \propto m^{-\alpha_i}, \tag{1.10}$$

$$\alpha_i = \begin{cases} 0.3 \pm 0.7, & 0.01 \le m/M_\odot < 0.08, \\ 1.3 \pm 0.5, & 0.08 \le m/M_\odot < 0.50, \\ 2.3 \pm 0.3, & 0.50 \le m/M_\odot < 1.00, \\ 2.3 \pm 0.7, & 1.00 \le m/M_\odot. \end{cases} \tag{1.11}$$

and $\xi(m)\,\mathrm{d}m$ is the number of single stars in the mass interval m to $m+\mathrm{d}m$. We can see, then, that the slope of the IMF varies, and the uncertainty is indeed very high in some intervals. At first glance, it might seem that the Kroupa IMF has three breaks. However, note that at $1\,M_\odot$, only the quoted uncertainty of the slope changes, not the slope itself. Therefore, in our simulations, we use a single slope of $\alpha = 2.3$ for the entire interval $m > 0.5\,M_\odot$. For practical purposes, the IMF is often expressed in logarithmic form,

$$\xi_L(\log_{10} m) = \ln(10)\, m\, \xi(m), \tag{1.12}$$

which describes the number of stars per logarithmic mass interval.

It is also worth noting that different mass ranges contribute differently to the total stellar population and mass budget. Approximately half of the total mass is contained in stars with $0.01 \le m \le 1\ M_\odot$, and the other half in stars with $1\text{–}50\,M_\odot$. If we also take stellar physics into account, the IMF suggests that, although low-mass stars are the most numerous, the properties of stellar populations are largely determined by the most massive stars.

The Kroupa IMF is shown as the yellow curve in Figure 1.1. Note the two breaks mentioned and their corresponding slopes; for lower-mass stars, the slope is gentler, reflecting their abundance, while for more massive stars, the slope is steeper, and their abundance thus declines rapidly.

1.1.4 Chabrier IMF

The last form of the IMF that we will discuss in this thesis is the Chabrier IMF, which was formulated by Gilles Chabrier in a 2003 paper [8]. He based his work on previous studies and focused primarily on the IMF as it relates to low-mass stars and brown dwarfs.

Chabrier examined the issue of unresolved binary or multiple star systems in greater detail and made the necessary corrections. He also paid attention to brown dwarfs. Although these objects do not contribute significantly to the total mass, they are numerous and influence the shape of the IMF in the low-mass region. He also compared the IMF of the galactic disk, young star clusters, galactic spheroids, globular clusters, and the dark halo and early stars. He observed how its form changes and whether this change is significant. When we take into account observational biases and the dynamic evolution of systems, it appears that the IMF has a very similar shape across different environments, which supports the idea that the IMF is nearly universal, at least in the local universe. (We will discuss the question of universality in more detail in a later section.)

Based on observations of the Galactic Disk, Chabrier proposed a modern single-object IMF parameterization that combines a log-normal shape at low masses with a power-law relationship at higher masses:

$$\xi(\log m) = A \exp\left[-\frac{(\log m - \log m_c)^2}{2\sigma^2}\right], \quad m \lesssim 1\,M_\odot. \tag{1.13}$$

$$\xi(m) \propto m^{-\alpha}, \quad m \gtrsim 1\,M_\odot. \tag{1.14}$$

He specified the normalization constant as $A = 0.158$, the standard deviation as $\sigma = 0.69$, and the characteristic mass as $m_c = 0.079\,M_\odot$. The paper [8] also acknowledges the relevant uncertainties and the models used to arrive at these values.

Thus the IMF can be roughly divided into two parts with a transition point around $1\,M_\odot$, which, based on Chabrier's interpretation, can be understood as follows: For low masses, there is a characteristic mass around which the largest number of stars form. Most stars, therefore, do not form uniformly across the entire mass range, but are concentrated around a typical value. This shape is usually associated with turbulence in molecular clouds, which creates structures of varying density and leads to the formation of stellar cores of different masses.

At higher masses, the IMF transitions to a power-law distribution, similar to the original Salpeter function. In this region, there is no longer a single characteristic mass, and the distribution becomes approximately scale-free; the number of stars decreases continuously with increasing mass. This shape suggests that processes associated with gravitational collapse and accretion play an important role in the formation of more massive stars, although they do not fully explain their formation. Chabrier thus provided an important

empirical parameterization of the IMF and discussed possible physical interpretations of its shape, particularly for the low-mass log-normal part.

As in the previous cases, the shape of the Chabrier IMF is shown in Figure 1.1. We see a smooth curve that at $1\,M_{\odot}$ transitions into a straight line with the appropriate slope.

1.1.5 What is more than IMF?

In recent years, the classical IMF has been extended to larger scales. From the level of individual star clusters, the IMF concept has been expanded to the level of entire galaxies in the form of the so-called integrated galactic initial mass function (IGIMF). As formulated by Weidner & Kroupa (2006) [9] and further explored by Haas & Anders (2010) [10], this framework is based on the assumption that stars form predominantly in star clusters; therefore, the overall distribution of stars in a galaxy is not determined by a single universal IMF, but arises as the integral contribution of the IMFs of individual star clusters. A key role here is played by the distribution of star cluster masses, known as the cluster mass function (CMF), which determines how frequently star clusters of various masses form, and its shape resembles the power-law form of the classical IMF.

Within the IGIMF, the star formation rate (SFR) plays a significant role and is considered the primary factor in the formation of the global IMF. As summarized by Bastian et al. (2010) [11], in galaxies with low star formation rates, the low SFR statistically limits the maximum mass of the embedded star clusters, and thus the probability of massive stars forming also decreases. This then raises the question of whether the IGIMF is steeper than the standard Kroupa or Salpeter functions for individual star clusters. Conversely, in galaxies with high star formation rates, more massive star clusters can form, and therefore more massive stars become more likely. This may lead to a flatter IGIMF, which affects the overall evolution of the galaxy, its chemical composition, and so on.

1.2 Any problems, the IMF?

Over the years, we have gradually refined our understanding of the IMF, yet many questions related to it remain unanswered. Key research questions include whether the IMF is universal, its limits, its evolution, its variability, and its realization... We will touch on these questions briefly to give the reader a sense of the scope of the entire IMF situation.

1.2.1 Limits of the IMF

The IMF is not defined for stars of all masses, but there are upper and lower limits to this function. The lower limit is generally considered to be approximately $0.01\,M_{\odot}$; below this value, we transition to objects of a planetary nature. The range from 0.01 to $0.08\,M_{\odot}$ includes brown dwarfs, although there is ongoing debate as to whether they should be included in the IMF. Consequently, different authors treat them differently: for example, Kroupa adjusts the slope of the IMF for this region (equation (1.11)) or separates it from the stellar part, while Chabrier naturally includes brown dwarfs in the IMF and emphasizes their significance (he uses only one threshold $1\,M_{\odot}$).

The upper limit is often given as approximately $150\,M_{\odot}$, although its exact value remains a subject of debate. While the IMF allows for the existence of stars with masses greater than $150\,M_{\odot}$, their absence calls this assumption into question. Donald F. Figer in his work [12] studied the Arches star cluster, which is both massive enough to form extremely heavy stars and young enough that these stars have not yet had time to die out as supernovae. Although a certain number of extremely massive stars were expected based on an extrapolation of the IMF, in reality, no stars with masses greater than $130\,M_{\odot}$ have been observed.

To conclude, note that the IMF covers a range of about four orders of magnitude.

1.2.2 Universality and Variability of the IMF

The universality of the IMF has been studied by numerous astrophysicists; for example, Pavel Kroupa and Gilles Chabrier, mentioned earlier, discuss this issue in their works. In this context, we should again mention a more recent review study by Nate Bastian [11], which addresses this question systematically.

The authors generally lean toward the view that the IMF is approximately universal. Although certain differences in its shape appear when observing various stellar systems, these deviations can in most cases be explained without the need to assume actual physical variations in the IMF.

One of the main sources of uncertainty is the fact that the IMF is derived from the luminosity function. The conversion between observable quantities and stellar mass depends on stellar evolution models, which themselves contain significant uncertainties. Another important factor is selection effects. For example, in young star clusters, brighter – and thus typically more massive – stars are more easily detectable than faint stars, which can lead to systematic bias in the derived IMF.

Multiple star systems also play a significant role, as they may be observed as a single object, thereby further influencing the resulting shape of the IMF. The authors therefore note that differences between individual studies often do not reflect actual physical changes in the IMF, but rather differences in analytical methods and the models used.

Physical conditions of the environment, such as metallicity, gas density, and temperature, or possibly the level of turbulence, could argue against the universality of the IMF. These factors influence the fragmentation process of molecular clouds and, consequently, the resulting distribution of stellar masses. Differences in star formation in extreme environments also tend to support the idea of IMF variability. Examples of such environments in the Milky Way include so-called *starburst clusters*, which are characterized by stellar densities several orders of magnitude higher than those of ordinary nearby star clusters. Based on this, a version of the IMF shifted toward higher masses, referred to as a *top-heavy IMF*, is considered, in which the relative proportion of more massive stars is higher than in the canonical form. Giant elliptical galaxies present a similar contradiction of the opposite nature; based on them, scientists are instead considering a so-called *bottom-heavy IMF*, meaning there are more low-mass stars than the canonical IMF would predict.

Although the data in many cases explain the observed variability and thus support the idea of the universality of the IMF, there is insufficient evidence to rule out either possibility. More precise and extensive observations in the future should make it possible

to answer this question.

1.2.3 Stellar populations and the IMF

Nothing remains unchanged over time, and this is also true for the IMF. As the universe evolved and took shape, the physical conditions within it also changed. Stars that formed in the early universe differed from stars forming today. Population III stars, as the first stars, formed from pristine gas with zero metallicity. Based on this, theoretical models generally suggest that these stars may have been, on average, more massive and closer to the previously mentioned top-heavy IMF (e.g. [11]). Younger stars formed in environments increasingly enriched with metals, so their IMF is closer to the more commonly used canonical form.

1.2.4 Origin of the IMF

Another question that remains unanswered is the physical origin of the IMF. While we can now describe its shape with reasonable accuracy, the question of why it takes this particular form still remains. Several explanations have been proposed, the most significant of which are models of turbulent fragmentation of molecular clouds, as discussed, for example, by Chabrier [8], who addresses this question only in the context of low-mass stars.

Other physical processes considered to influence the IMF include gravitational collapse, the thermal physics of gas and the properties of dust, feedback from forming stars, magnetic fields, and accretion of matter. The shape of the IMF is therefore likely the result of a combination of several physical processes, rather than a single mechanism, as discussed in [13]. He concludes that different processes operate differently in different parts of the mass spectrum. Gravitational collapse and turbulence are generally considered important to the formation of more massive stars, while for less massive stars, gas thermal physics and processes associated with fragmentation play a particularly important role.

1.2.5 Realization of the IMF

So far, we have discussed the IMF only from a theoretical perspective. However, the IMF is not only a mathematical function, but also a practical tool used to describe real stellar populations. Although it is defined as a continuous distribution of stellar masses, stars themselves form as discrete objects. This raises a fundamental question: how can a continuous function be used to generate a finite set of stars? The answer is not unique. Several approaches have been developed to interpret and implement the IMF in practice, each based on different assumptions and leading to different outcomes.

There are two approaches that represent, in a sense, the two limiting ways of thinking: *random* and *optimal sampling*. Random sampling is historically much older than optimal sampling, much more widely used, and more widely accepted. Naturally, it creates stochastic scatter between individual realizations, which can, however, differ significantly from one another, especially in the region of low-mass star clusters. Optimal sampling represents a newer approach. It is based on a fundamentally different interpretation of the IMF than random sampling; namely, it is a purely deterministic approach with no Poisson

noise. Optimal sampling also leads to a clear relation between $m_{\max}$ and $M_{\rm cl}$, which in random sampling exists only in a statistical sense and can be masked by the wide dispersion of individual realizations. In the following two chapters, we will examine both approaches in more detail to provide the reader with a thorough understanding of their assumptions, properties, and implications for modeling stellar populations.

Chapter 2

How Random is Random Sampling?

In Chapter 1, we provided a general overview of the IMF. In this chapter, we will take a closer look at one way of understanding and working with the IMF, specifically in the context of random sampling.

In its simplest and most basic statistical interpretation, the IMF is understood as a probability density function; that is, each star is randomly selected from the IMF, independently of the others. Although the IMF itself is defined as a continuous function, real stellar populations arise as discrete realizations of this underlying distribution. In mathematical terms, we can express this probability as follows:

$$p(m) = \frac{\xi(m)}{\int_{m_L}^{m^*_{max}} \xi(m)\,\mathrm{d}m}, \tag{2.1}$$

where $\xi(m)\,\mathrm{d}m$ represents the number of stars in the mass interval (note the emphasis on the word "number"), and normalization is performed over the adopted mass range.

Normalization is essential for the complete implementation of the IMF as a probability density function; it ensures that the total area under the curve is equal to unity within the limits we define. Although the Kroupa IMF is often written down to $0.01\,M_\odot$, in our simulations we use only the stellar mass range $0.08 \leq m/M_\odot \leq 150$, excluding objects below the hydrogen-burning limit. This corresponds to the fundamental axiom of probability that the random event in question will certainly occur within the given interval. In the case of multiple power-law systems, such as the Kroupa IMF we are currently using, normalization additionally enforces continuity throughout the entire domain at the break points, specifically $0.08\,M_\odot$ and $0.5\,M_\odot$ (i.e., the left and right limits are equal). On the other hand, the function is not differentiable at the break points, which is a consequence of its piecewise definition and the price we pay for its relative simplicity.

IMF sampling can be based on two parameters, each of which has different implications for the model. In the first case, the determining parameter is the number of stars in the star cluster; in the second case, the limiting condition is the mass of the star cluster, $M_{\rm cl}$. The probability introduced in equation (2.1) refers to the number of stars in a given mass interval and is stated without any global constraint. Therefore, if we wanted to perform sampling directly in accordance with this probability, we would specify a fixed number of stars, N. In that case, the total mass $M_{\rm cl}$ would be an unknown outcome of the random selection. Fixing N is a perfectly valid and classical approach from a mathematical point of view.

However, the first approach does not make much physical sense and is a bit unrealistic, since a star cluster forms from a finite mass reservoir rather than from a predetermined number of stars. When the sampling is constrained by total mass, we are dealing with a conditional sampling problem, which has properties different from the original unconstrained form. The independence of individual draws is partially reduced, and correlations may arise among the properties of the resulting stellar population. We will address this issue in more detail in Section 2.2, but first we will introduce the random sampling procedure we have chosen for this work, on the basis of which we will perform future simulations.

2.1 Implementation

In order to properly analyze random sampling, we needed to create a usable code. Implementing random sampling is not particularly difficult; one simply needs to select a preferred IMF (in our case, the Kroupa IMF) and use it to randomly generate stars until the desired number or mass is reached.

Despite the simplicity of the principle itself, choosing a specific algorithm is not entirely trivial. It is important to bear in mind that the method used to terminate the selection process can significantly affect the properties of the simulated star clusters. The influence of different IMF sampling methods and stopping conditions, including stop-before, stop-after, and stop-nearest prescriptions, was investigated, for example, by Haas & Anders (2010) [10], who showed that these choices can have a substantial impact on the resulting properties of stellar populations.

With this in mind, our implementation was only slightly more complicated: from a physical standpoint, we chose the total mass $M_{\rm cl}$ as a fixed input parameter. We then gradually populated the star cluster with stars whose masses were randomly selected from the IMF. We repeated this process until the total mass of the star cluster was reached (within a given tolerance). If a newly drawn star caused the target mass to be exceeded, the draw was rejected and the selection was repeated. We therefore interpret the following quantitative results in the context of this particular stopping and rejection prescription.

We view this procedure primarily as a simple statistical method for implementing a finite-mass constraint; it is not a detailed description of the actual star formation process, where gas fragmentation, accretion, feedback, and dynamical interactions also play a role. As mentioned, introducing a fixed finite $M_{\rm cl}$ may violate the independence of individual draws, which is further strengthened by our algorithm by discarding stars that are too massive.

Apart from the total mass, our algorithm imposes no further constraints on the internal structure of the resulting star cluster. However, observations have motivated a long-standing discussion about a possible relation $m_{\rm max}$–$M_{\rm cl}$, which was explored, for example, by Weidner and Kroupa (2006) [9], who introduced the so-called sorted sampling method. Stars are randomly selected from the IMF, sorted by mass, and added to the star cluster so that the resulting population better matches the observed relation $m_{\rm max}$–$M_{\rm cl}$. This approach physically mimics a scenario where star formation proceeds until feedback from the most massive stars halts the process. Our algorithm does not prescribe any explicit relation between $m_{\rm max}$ and $M_{\rm cl}$; this relation emerges only statistically as a consequence of random selection and the constraint on total mass, as we will discuss later in Section 2.2.2.

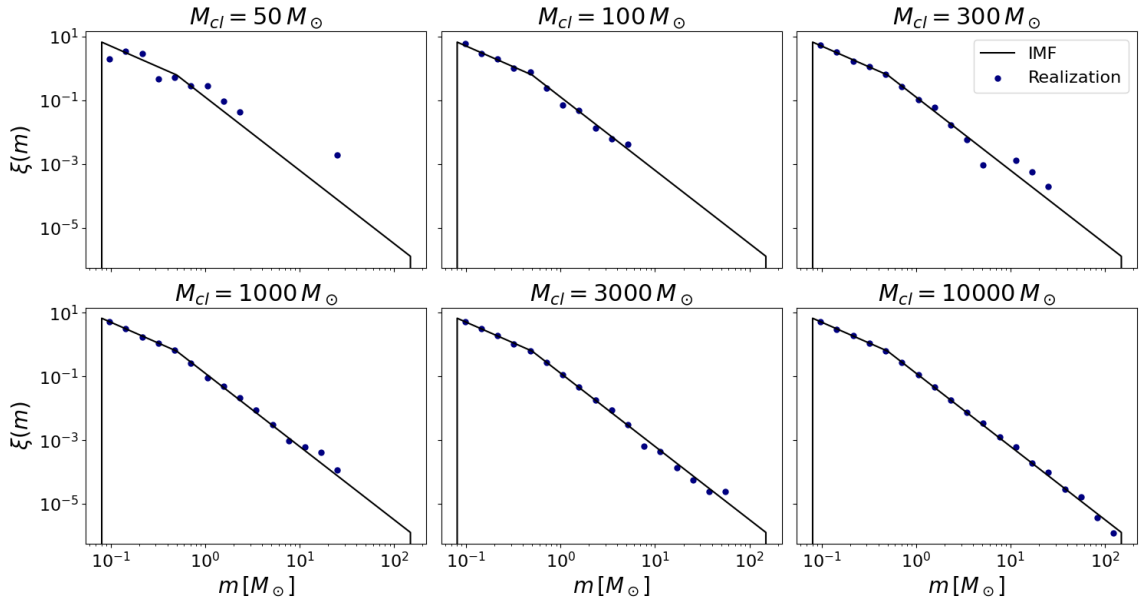


Figure 2.1: Realizations of random sampling for star clusters with different total masses (50–10000 $M_\odot$). Each panel shows one stochastic realization compared to the underlying IMF. The plotted points represent binned number densities computed from the sampled stellar masses in logarithmic mass bins and normalized for comparison with the continuous IMF curve. For low-mass star clusters, the sampled data do not cover the full theoretical mass range and exhibit larger fluctuations around the IMF curve.

Simulations with a limited stellar mass reservoir also appear in other studies. For example, Applebaum et al. (2020) [14] used discrete stochastic IMF sampling in simulations of dwarf galaxies instead of assuming a continuously populated IMF. They showed that, in small systems, the finite available mass can significantly influence how the IMF is actually populated, particularly at the high-mass end.

We started with just a general overview. Figure 2.1 illustrates the behavior of random sampling for star clusters with different total masses. We considered six values 50, 100, 300, 1000, 3000, 10000 $M_\odot$ and plotted one realization of randomly selected stars for each of them. For star clusters with low total mass, naturally, the entire IMF is not covered (such a realization contains only a limited number of stars), and in particular, the most massive stars often do not appear at all. The dispersion is also significant in the region of massive stars, where individual realizations always follow the theoretical IMF curve with a certain deviation.

As the following Figure 2.2 shows, individual simulations differ from each other, which is particularly noticeable for low-mass star clusters. The stochastic dispersion is most pronounced for massive stars, since there are few of them in the cluster. In contrast, for a star cluster with a mass of 10000 $M_\odot$, the simulations closely follow the theoretical IMF, with deviations appearing only at the high-mass end, since even here the proportion of massive stars is significantly smaller compared to less massive ones.

Figure 2.3 also illustrates the variation in dispersion for different cluster masses. Here, we used the Kolmogorov–Smirnov test to quantify the difference between individual real-

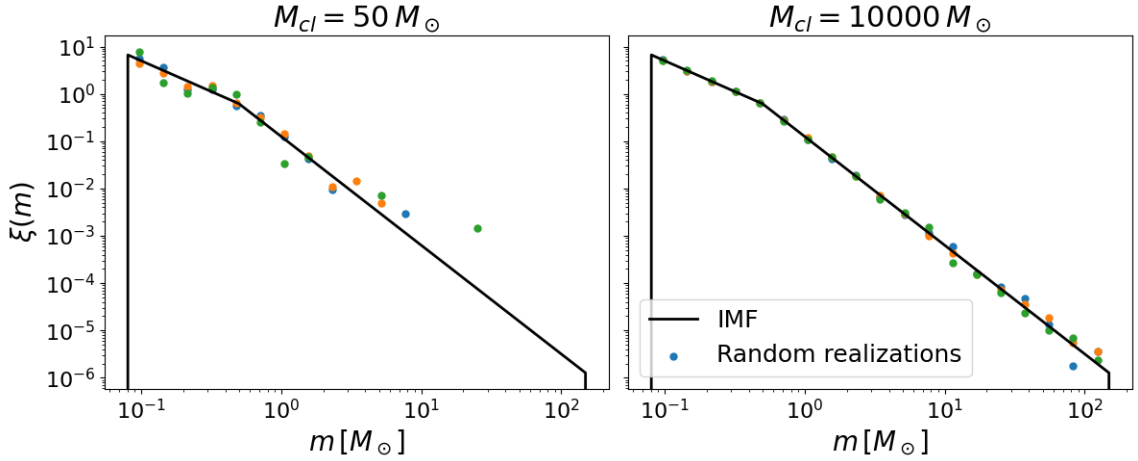


Figure 2.2: Realizations using random sampling for star clusters with masses of $50\,M_\odot$ and $10000\,M_\odot$. Star clusters with $50\,M_\odot$ show larger deviations from the theoretical curve and do not cover the full mass range; in contrast, for clusters with $10000\,M_\odot$, deviations are confined mainly to the high-mass end.

izations and the theoretical IMF in order to assess how strongly these realizations deviate from the given distribution. For each value of $M_{\rm cl}$, we generated 1000 random realizations. The K–S distance was computed from the unbinned stellar masses in each realization and compared with the theoretical cumulative distribution function of the adopted IMF. As we can see, with increasing cluster mass, not only does the total deviation decrease but so does the variance of the realizations; that is, for massive clusters, random sampling more accurately reproduces the theoretical IMF. Less massive clusters are more affected by stochasticity, and thus both the deviation and the variance are significantly higher.

2.2 The Four-Leaf Clover of Assumptions

Although the IMF is considered a standard probability density function, several assumptions must be taken into account when using random sampling. First, it is assumed that the star masses are selected independently; that is, the formation of one massive star does not prevent the formation of another massive star. Second, it is assumed that all stars are selected from the same underlying IMF. Third, random sampling imposes no global constraints on the stellar population, which in principle allows for the formation of a very massive star even in a low-mass stellar system. Finally, this approach is inherently stochastic, meaning that different realizations of the stellar population can vary significantly.

In this context, a stellar system can be understood as a set of independent samples drawn from the IMF, provided that the selection is based on a fixed number of stars N. In this case, the assumptions are entirely valid. In practice, however, it makes more sense to model stellar populations based on $M_{\rm cl}$; in this case, the assumptions are not entirely valid, as correlations are introduced between the variables, leading to changes in the overall statistical properties of the system.

In the following subsections, we will examine each assumption in greater detail. We

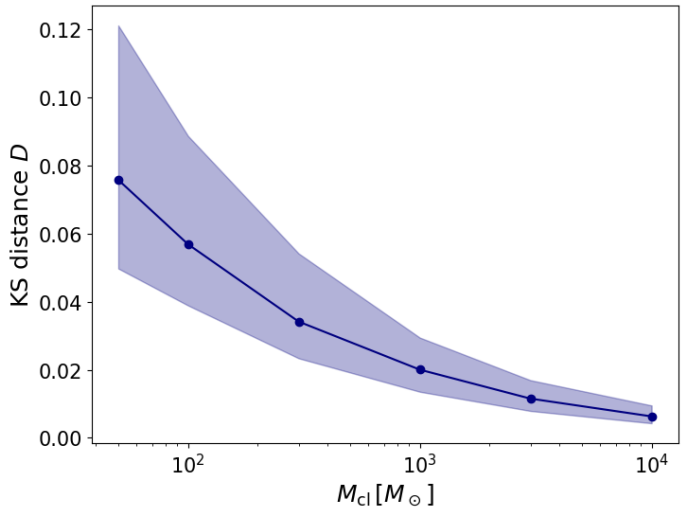


Figure 2.3: The Kolmogorov–Smirnov distance D between realizations and the theoretical IMF as a function of $M_{\rm cl}$. The points represent the median values; the shaded region corresponds to the 10th–90th percentile. As the cluster mass increases, both the deviation and its variance decrease.

will focus primarily on the effect of $M_{\rm cl}$, the newly established correlations, and, more generally, on the consequences resulting from the imposed mass limit. Note that our algorithm amplifies these correlations and that the observed relationships do not arise solely as a direct effect of the mass limit.

2.2.1 The Assumption of Universality

We will discuss the universality assumption very briefly. In this context, the universality assumption is slightly different from the broader question of IMF universality discussed in Chapter 1. Random sampling does not in itself imply that the IMF is universal across the entire Universe; in principle, the IMF could vary with metallicity or other environmental or dynamical conditions. In our implementation, however, all star clusters and all realizations are generated from the same adopted IMF. In this sense, the IMF is treated as universal within the given sampling procedure. If the IMF were not universal, then our implementation of random sampling would represent a simplification that could lead to discrepancies when applied to real stellar populations.

2.2.2 The Assumption of Independence

First, we will consider the assumption that star masses are drawn independently from the IMF, meaning that the formation of one star does not influence the mass of another star, and each star represents an independent random draw from the given distribution. Based on this assumption, a stellar population can be interpreted as a set of independent realizations of a single probability function.

However, our algorithm yields results that do not fully support this assumption. One of them is illustrated in Figure 2.4, which shows the relation between the maximum stellar

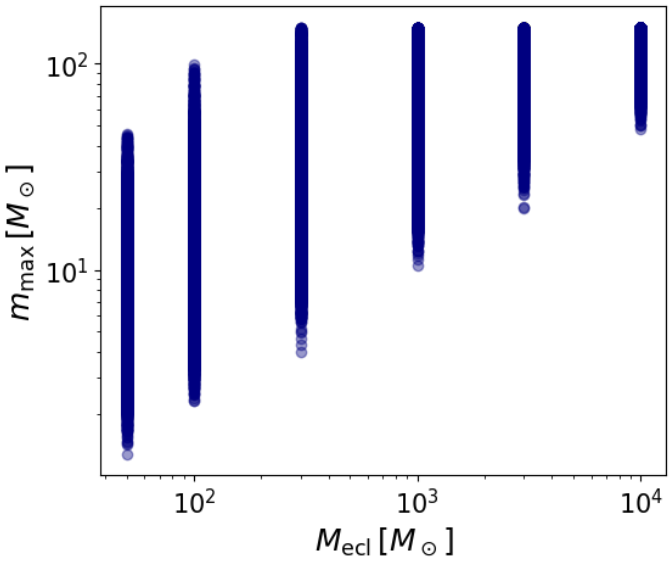


Figure 2.4: Relation between the maximum stellar mass $m_{\max}$ and the total mass of the star cluster $M_{\rm cl}$. For each of these masses, 10000 simulations were performed.

mass $m_{\max}$ and the total cluster mass $M_{\rm cl}$ for six different values of $M_{\rm cl}$. For each of these masses, 10000 simulations were performed. The variation across all runs is indeed large, especially for the first three cluster masses, where individual realizations can differ significantly from one another. Although we select stars from a mass range up to $150\,M_\odot$, $m_{\max}$ is influenced by the specified $M_{\rm cl}$; thus, for star clusters of $50\,M_\odot$ and $100\,M_\odot$, it is impossible to reach this physical upper limit. This is visible especially for $M_{\rm cl} = 100\,M_\odot$, where, in a small (but non-negligible) number of cases, the most massive star contains a substantial fraction of the total cluster mass. As the mass increases, $m_{\max}$ reaches the upper mass limit of $150\,M_\odot$ much more frequently, since the cluster mass is large enough to allow such stars to be sampled. In the range of very massive star clusters, the probability of occupying the highest part of the IMF would gradually approach 100%. The relation between $m_{\max}$ and $M_{\rm cl}$ is thus evident here, although it is not deterministic; rather, it is broad, stochastic, and characterized by substantial scatter.

If we continue further, we encounter a fact that does not exactly "play into the hands" of independence. Stars are selected independently from the theoretical IMF, but the condition imposing a limit on the total mass of the star cluster strongly influences this independence. A simple example is the relation between $m_{\max}$ and the number of stars in a star cluster N. We worked with the same set of values as for the graph in Figure 2.4, and we compared these values with cases in which the condition for the total $M_{\rm cl}$ was not imposed. In the unconstrained comparison sample, we did not impose any condition on the total cluster mass. Instead, for each mass-constrained realization, we kept the same value of N and generated the corresponding maximum stellar mass expected from N independent draws from the IMF. In the unconstrained case, the listed values of $M_{\rm cl}$ served only as labels for the corresponding mass-constrained samples from which we took the values of N.

For clarity, we present only three representative examples corresponding to cluster masses of 50, 1000 and $10000\,M_\odot$, in order, among other things, to compare how this

effect changes with increasing cluster mass. A graphical comparison is shown in Figure 2.5; the numerical values indicating the resulting dependence are listed in Table 2.1.

Table 2.1: Spearman's correlation coefficient ρ and p-values for the relation between the maximum stellar mass $m_{\max}$ and the number of stars N in a star cluster. A strong negative correlation arises only when restricted to the total mass of the star cluster, while in the unrestricted case the relation is weak.

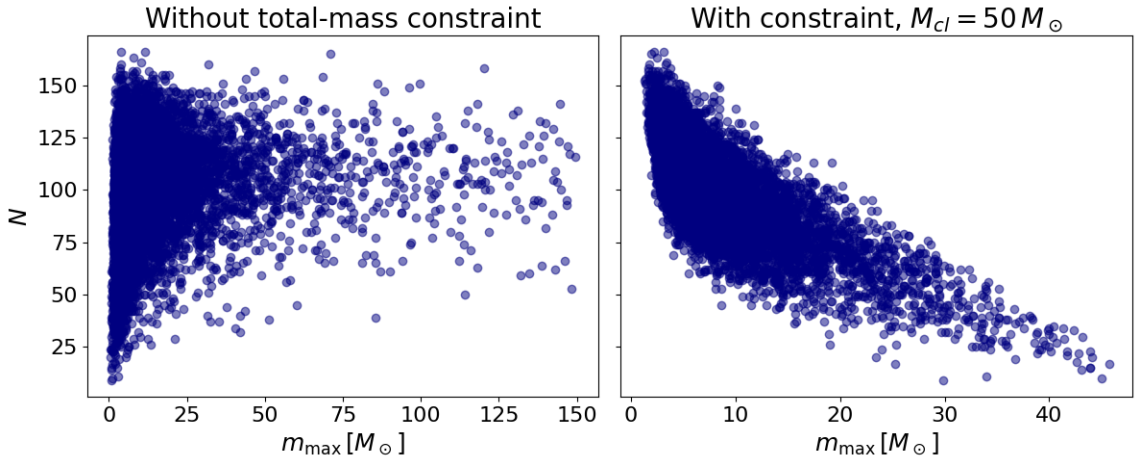
	With constraint		Without constraint	
$M_{\rm cl}$ [$M_\odot$]	ρ	p-value	ρ	p-value
50	-0.751	$p \ll 10^{-10}$	0.193	$p \ll 10^{-10}$
1000	-0.671	$p \ll 10^{-10}$	0.078	$p \ll 10^{-10}$
10000	-0.340	$p \ll 10^{-10}$	0.045	$p < 10^{-5}$

This dependence is most pronounced for low-mass star clusters. The correlation here is indeed high, and the low p-value confirms that it is not caused by random fluctuations. As the total cluster mass increases, the correlation weakens, the negative trend becomes less distinct, and the differences between the constrained and unconstrained cases diminish. In our simulation, this effect arises from the fixed constraint $M_{\rm cl}$: if a very massive star is drawn, a smaller amount of mass remains available for the rest of the population, leading to a smaller total number of stars. In more massive star clusters, however, the total mass reservoir is sufficiently large that even a very massive star represents only a smaller fraction of the total mass. As a result, both massive and low-mass stars can form in larger numbers.

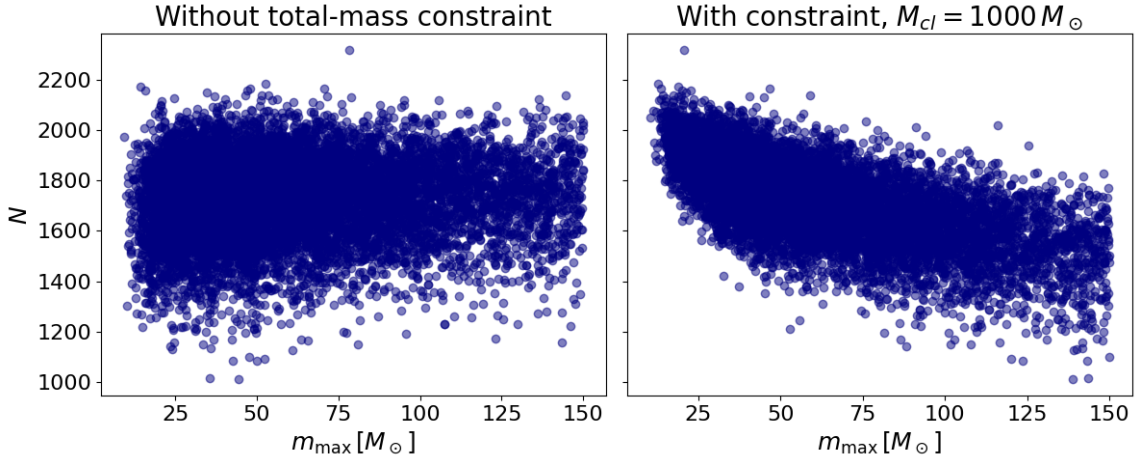
In contrast, in the unconstrained case, no strong dependence between $m_{\max}$ and N can be identified. The distribution of points does not show an obvious structure and Spearman's correlation coefficient ρ remains small in all three cases. The realizations here are influenced only by the number of stars N; the probability of selecting a massive star increases with N, which is a natural consequence of the statistics of extreme values. Although some of the corresponding p-values are formally very low, this mainly reflects the large number of realizations rather than a physically strong correlation. This supports the interpretation of selection within random sampling and indicates that the observed dependence in the limited case does not arise from the properties of the IMF itself but is a consequence of the constraint on the total mass of the star cluster.

2.2.3 The Assumption of Absence of Global Limitations

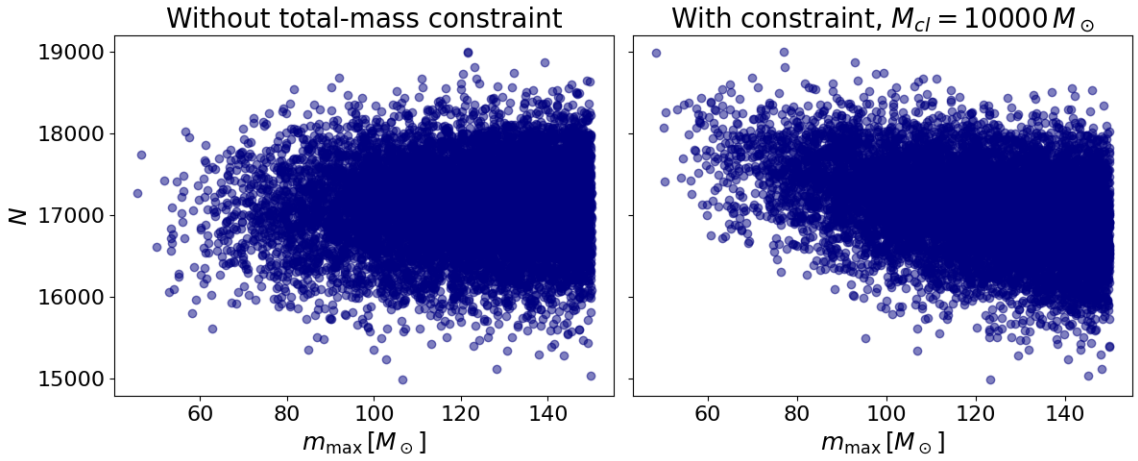
Another assumption, closely related to the previous one, is the absence of global limitations. The assumption of the absence of global limitations states that stars are selected from the IMF independently, without taking into account the overall properties of the system. This means, for example, that we do not impose a limit on the most massive star in a star cluster; therefore, even a low-mass star cluster can contain a very massive star. And here is the problem... The moment we impose a condition on the total mass, the assumption no longer holds. In this subsection, we will expand on the issue of independence; we will work with the relation between $m_{\max}$ and $M_{\rm cl}$, which we will then expand with additional observations.



(a)



(b)



(c)

Figure 2.5: The relation between the maximum stellar mass $m_{\max}$ and the total number of stars N for star clusters with different total masses $M_{\rm cl}$. For each mass, the simulation without a total mass limit is shown on the left, and the simulation with a limit is shown on the right. In the case of the simulation without a limiting mass, no visible dependence is apparent in the data, which contrasts with the case with a limit, where the correlation is pronounced but decreases with increasing mass.

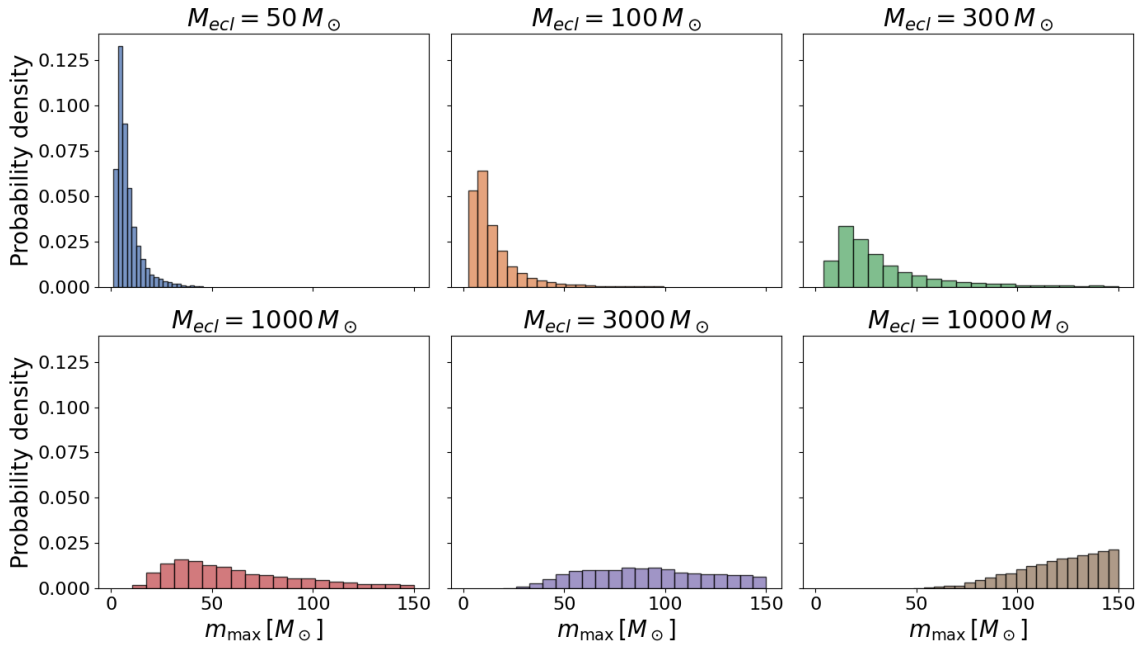


Figure 2.6: Probability density distribution of the maximum stellar mass $m_{\max}$ for star clusters with different cluster masses $M_{\rm cl}$. The individual panels correspond to fixed values of $M_{\rm cl}$ and show how, as the cluster mass increases, the distribution shifts toward higher values and simultaneously broadens.

Figure 2.6 shows the distribution of $m_{\max}$ for various values of the total mass of the star cluster $M_{\rm cl}$. Individual panels allow us to observe how this distribution changes as a function of the size of the system, while all other assumptions remain constant.

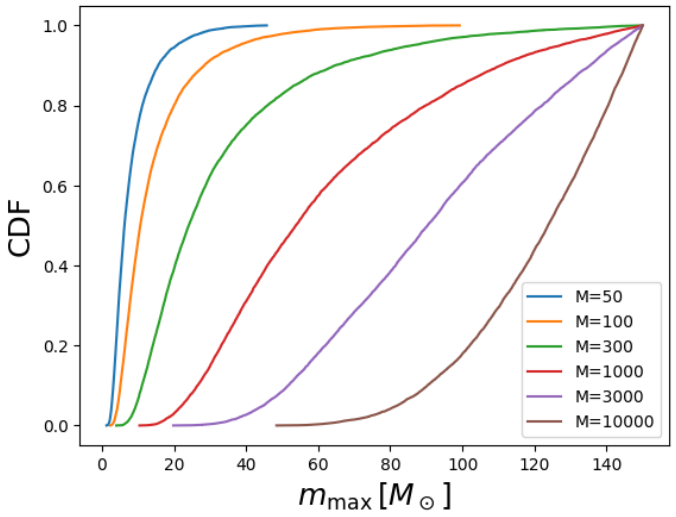


Figure 2.7: Cumulative distribution functions for star clusters of various masses. The curves illustrate the systematic shift toward higher masses with increasing cluster mass.

For low-mass star clusters, the distribution is concentrated mainly around low values; as

mass increases, the distribution widens, the probability of massive stars occurring rises, and the dispersion of the distribution also increases. More massive systems exhibit increased variability in the maximum mass of the most massive star. For the most massive star clusters, the distribution of $m_{\max}$ becomes affected by the adopted physical upper stellar mass limit $m^{*}_{\max} = 150\,M_{\odot}$. As a result, the distribution is truncated at the high-mass end and starts to accumulate near this upper limit. The occurrence of very massive stars is much more common, which is in accordance with the results of the relation between $m_{\max}$ and N. The maximum stellar mass $m_{\max}$ thus behaves as a stochastic quantity, where its value is again influenced by the total mass and the specific realization. If we were to consider star clusters with even higher masses, the concentration of the distribution at its right end would be even more pronounced.

We further supplement these conclusions with the cumulative distribution function (CDF) shown in Figure 2.7, which indicates the probability that the value of $m_{\max}$ is less than or equal to a given value, again across star clusters of various masses. The curves systematically shift toward higher masses until they again "hit" the upper mass limit of $150\,M_{\odot}$. As before, we see that the probability of higher maximum masses increases with the total mass of the star cluster.

To conclude this subsection, we will add one more piece of information: we set a threshold of $20\,M_{\odot}$ and tracked the probability of a star of this mass occurring across star clusters (see Table 2.2). For star clusters of $50\,M_{\odot}$, this probability is very low (about 5.5%), yet it is not entirely negligible. Conversely, for star clusters with masses of $1000\,M_{\odot}$ and greater, the occurrence of such a massive star is already almost certain or certain.

Table 2.2: Probability that at least one star more massive than $20\,M_{\odot}$ occurs, as a function of the total mass of the star cluster $M_{\rm cl}$.

$M_{\rm cl}$ $[M_{\odot}]$	$P(m_{\max} > 20\,M_{\odot})$
50	0.055
100	0.201
300	0.609
1000	0.970
3000	1.000
10000	1.000

For low-mass star clusters (considering 50–300 $M_{\odot}$), the $M_{\rm cl}$ constraint also slightly alters the shape of the IMF, since the most massive stars are either extremely rare or absent altogether. This is illustrated in Figure 2.8, where the data are divided into bins and we examine the number of stars in each bin. Table 2.3 provides a more detailed description of the individual bins. For low-mass stellar clusters, the number of massive stars drops sharply, and the IMF deviates from the theoretical curve.

2.2.4 The Assumption of Stochasticity

Finally, we turn to the assumption of stochasticity, whose effects have already been observed in the previous sections. This assumption implies that individual realizations can differ

significantly, even if they share the same global properties, such as the total mass of the star cluster.

In particular, the occurrence of massive stars is governed by probability, which naturally leads to variance in observable quantities, as reflected in the relations between $m_{\max}$ and $M_{\rm cl}$ and between $m_{\max}$ and N. This also means that the theoretically predicted IMF

Table 2.3: Mass bins used in the analysis together with their approximate typical stellar classes. Bins are divided into the range 0.08–150$M_\odot$ using a logarithmic scale so that the stars are distributed more evenly across the intervals.

	Mass range [$M_\odot$]	Typical class
Bin 1	0.08 – 0.17	M
Bin 2	0.17 – 0.36	M
Bin 3	0.36 – 0.77	M / K
Bin 4	0.77 – 1.63	G / F
Bin 5	1.63 – 3.46	F / A
Bin 6	3.46 – 7.36	B
Bin 7	7.36 – 15.64	B / O
Bin 8	15.64 – 33.23	O
Bin 9	33.23 – 70.60	O
Bin 10	70.60 – 150.00	O / WR

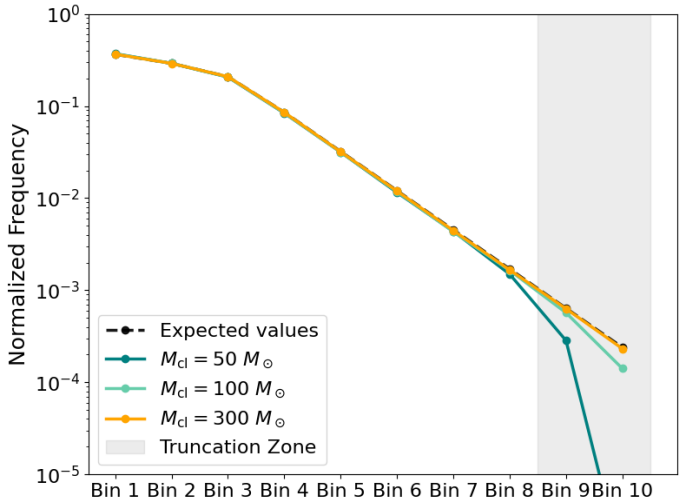


Figure 2.8: The effect of the total $M_{\rm cl}$ constraint on the discrete distribution of stars in mass bins. The black dashed line represents the expected theoretical values, calculated as the average distribution of a reference star cluster with a mass of 10000$M_\odot$, while the colors indicate the mass distribution in individual bins for star clusters of varying masses. As mass increases, the curves approach the theoretical values. The gray "truncation zone" highlights the high-mass region, where the probability of stars occurring is significantly suppressed by the introduction of $M_{\rm cl}$.

profiles can only be achieved in the limit of large star clusters, whereas in smaller ones,

statistical fluctuations will always occur. As in the case of the previous assumptions, it is important to keep in mind that, in practice, this stochasticity is partially constrained by the total cluster mass. This constraint reduces the range of possible realizations and introduces dependencies between individual stars, although this effect is less pronounced than in the previous cases.

Although these four assumptions provide the necessary mathematical framework for modeling the IMF as a continuous probability density, in real systems, we encounter their physical limits. It turns out that the purely stochastic selection model is in direct conflict with the existence of global constraints, particularly the finite mass of the parent star cluster $M_{\rm cl}$. Once we set the implicit limit of the $M_{\rm cl}$ condition to $m_{\rm max}$, the generated realizations from the IMF no longer strictly satisfy the ideal assumption of independent draws and thus become conditional on the imposed global constraint.

However, although the $M_{\rm cl}$ constraint does indeed have a significant impact on the properties of a star cluster, this impact is not equally important in all regimes. As the mass of the star cluster increases, this effect weakens, since the mass of a single star accounts for an increasingly smaller fraction of the total mass reservoir. The properties of more massive star clusters are therefore no longer as sensitive to individual draws, as we have indeed seen in our simulations. The differences between random sampling with a fixed N and with a fixed $M_{\rm cl}$ thus gradually diminish, and in the limit of very massive star clusters, the two approaches converge in a relative sense. For large star clusters, this violation of independence is therefore less of a concern than for low-mass systems; however, since small star clusters constitute a significant portion of the star cluster population, we should not overlook the consequences of the $M_{\rm cl}$ constraint.

Chapter 3

How Optimal is Optimal Sampling?

The second approach to sampling stars from the IMF, which we will discuss in this chapter, is known as optimal sampling. Although we interpret random sampling as a random and independent process, optimal sampling represents the exact opposite, in which the sampling contains no Poisson noise. Instead, the IMF acts here as a deterministic prescription for constructing a stellar population, meaning that for two star clusters with the same initial mass, it leads to the same distribution of stars. The masses of stars are therefore not random quantities but fixed values determined by the total mass of the system and the shape of the IMF itself.

This approach was introduced by Kroupa et al. (2013) [7], and its mathematical formulation was further refined and improved by Schulz et al. (2015) [15], among other reasons, due to doubts about random sampling, which exhibits significant stochastic fluctuations, particularly in low-mass star clusters, and does not always reliably reproduce the observed relationship between the maximum stellar mass $m_{\max}$ and the total cluster mass $M_{\rm cl}$. The approach assumes that stars influence one another and that their masses are determined to some extent by initial conditions such as temperature, rotation, magnetic fields, and so on. As a result, stellar masses are not fully independent, but may exhibit a certain degree of mutual dependence. In this framework, the IMF is no longer treated as a probability density function but rather as a deterministic distribution that can be directly translated into a discrete stellar population. Optimal sampling therefore discretizes the continuous IMF function into specific masses of individual stars, so that the resulting population perfectly represents the intended IMF shape without statistical noise, which, of course, has a major impact on the overall modeling of stellar populations. Observable quantities (ionization luminosity, number of massive stars, etc.) are thus fixed and become unambiguous functions of $M_{\rm cl}$.

To illustrate the mechanism of optimal sampling, we mainly used [7]. The mechanism of optimal sampling can be divided into several essential steps:

1. **Definition of the most massive star** $m_{\max}$

   Let us consider the function $\xi(m)$ as a continuous function that specifies the number of stars per unit mass interval. Given the importance of the relation between $m_{\max}$ and $M_{\rm cl}$, it is necessary to first determine the mass $m_{\max}$, which is calculated using an iterative solution based on two fundamental equations, namely the equation defining

the normalization condition

$$1 = \int_{m_{\max}}^{m_{\max *}} \xi(m)\,\mathrm{d}m, \tag{3.1}$$

where $m_{\max *}$ is the physical upper limit on the mass of stars, in our case $150\,M_{\odot}$. This condition ensures that exactly one star is expected above a certain mass, and thus determines the upper limit of the distribution. The second one is the closure condition (or mass-conservation condition), by which the total mass of the star cluster, after subtracting the mass of the most massive star, must equal the integral of the mass function from the lower limit m_{L} to $m_{\max}$

$$M_{\mathrm{cl}} - m_{\max} = \int_{m_{\mathrm{L}}}^{m_{\max}} m\xi(m)\,\mathrm{d}m. \tag{3.2}$$

2. Generating a sequence of stellar masses

After determining $m_{\max}$, we calculate the masses of the remaining stars by dividing the IMF into intervals of varying lengths, where each interval represents one star. The boundaries of these intervals are chosen so that

$$1 = \int_{m_{i+1}}^{m_i} \xi(m)\,\mathrm{d}m, \qquad m_{\mathrm{L}} \leq m_{i+1} < m_i, \qquad m_1 \equiv m_{\max}. \tag{3.3}$$

The mass assigned to the star in the interval $[m_{i+1}, m_i]$ is then given by

$$m_i^{\star} = \int_{m_{i+1}}^{m_i} m\,\xi(m)\,\mathrm{d}m. \tag{3.4}$$

Stars are selected in this way from the top down, moving from $m_{\max}$ toward m_{L}, until the stellar mass reservoir set by M_{cl} is exhausted. Thus, the continuous IMF is discretized into a deterministic sequence of individual stellar masses.

To simplify the calculation of $m_{\max}$, it is also possible to use an approximation

$$\log_{10}\left(\frac{m_{\max}}{M_{\odot}}\right) = 2.56\log_{10}\left(\frac{M_{\mathrm{cl}}}{M_{\odot}}\right)\left[3.82^{9.17} + \left(\log_{10}\left(\frac{M_{\mathrm{cl}}}{M_{\odot}}\right)\right)^{9.17}\right]^{-1/9.17} - 0.38. \tag{3.5}$$

This relationship can be used as an effective alternative to equations (3.1) and (3.2), as it allows for a quick estimation of an upper limit without losing numerical accuracy. Figure 3.1 shows an illustrative example of modeling a star cluster using optimal sampling. The star cluster has a target mass of $M_{\mathrm{cl}} = 50\,M_{\odot}$. We can see that the intervals are visibly wider for more massive stars than for less massive ones, where we can no longer distinguish the individual intervals from one another.

As can be seen, M_{cl} in optimal sampling acts as an input parameter that explicitly determines the resulting distribution of stars. In contrast, in random sampling – when considered in its simplest form, i.e., without any restriction on the total mass – it represents more of a final product, which is obtained as the sum of the masses of N randomly selected stars.

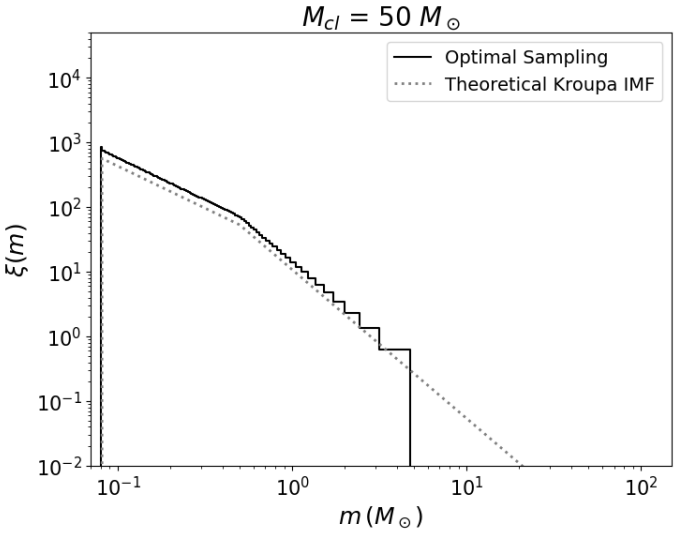


Figure 3.1: Comparison of the stellar distribution obtained using the optimal sampling method with the theoretical Kroupa IMF for a star cluster with target $M_{\rm cl} = 50\,M_\odot$ and total $M_{\rm cl} = 49.97\,M_\odot$. The most massive star has a mass of $m_{\rm max} = 4.75\,M_\odot$, and the star cluster contains a total of 113 stars. The solid line represents the discrete distribution, where the interval width varies, being widest for the most massive star in the cluster. The dashed line shows the theoretical curve.

A practical limitation of this numerical implementation of optimal sampling is that the specified cluster mass $M_{\rm cl}$ is not always reached exactly. In the final step of the algorithm, there is often at least a small amount of mass remaining from which it is no longer possible to form a complete star for the given integral; this results in what is known as residual mass. The total sum of the masses can therefore be slightly lower than the required $M_{\rm cl}$.

The optimal sampling method implemented in this way still had certain shortcomings. The main problem was the inability of the algorithm to accurately reproduce the total number of stars N in a way that also corresponded to the specified total $M_{\rm cl}$. For star clusters with a higher mass $m_{\rm max}$ ($> 100\,M_\odot$), this led to systematic deviations from the expected values derived from the theoretical integral. Therefore, Schulz et al. introduce a method of strict integral matching for each individual object in [15]. They do not treat the distribution as a whole, but define each i-th star as a separate object of interest. For each star, the integral over its mass interval must equal exactly one:

$$1 = \int_{m_{i+1}}^{m_i} \xi(m)\,\mathrm{d}m. \tag{3.6}$$

3.1 Implementation

Implementing optimal sampling is quite a bit more challenging than it was with random sampling. For instance, in addition to the inherent complexity, there is the added complexity of the multi-power-law form of the IMF, which requires taking into account different slopes for different mass intervals.

This time, we did not write this code ourselves, but based on the work of [16], where the code is publicly available, along with a detailed description. While their work focuses primarily on Galaxy-wide IMF (GalIMF), we adapted the relevant portion for our needs regarding IMF at the star cluster level. In general, GalIMF is a tool for calculating the IGIMF; it allows for modeling at the galactic level based on specified parameters such as star formation rates or the properties of star cluster populations.

For the calculation, we use the Kroupa IMF, which is defined slightly differently from the previous version (1.11)

$$\xi(m) = \begin{cases} k_1 m^{-\alpha_1}, & 0.08 \leq m < 0.5 \\ k_2 m^{-\alpha_2}, & 0.5 \leq m < 1.0 \\ k_3 m^{-\alpha_3}, & 1.0 \leq m \leq m_{\max}. \end{cases} \tag{3.7}$$

To ensure continuity throughout the range, the program must first solve the continuity conditions at the break points (0.5 and 1.0), just as it did for random sampling. We will express the overall system of equations in a simplified form, where we solve for a single unknown constant k_3, which we will determine based on normalization.

Since the function is piecewise, the total integral over the total mass breaks down into a sum

$$M_{\rm cl} = \int_{m_{\rm L}}^{m_{\rm s1}} m\xi_1(m)\,{\rm d}m + \int_{m_{\rm s1}}^{m_{\rm s2}} m\xi_2(m)\,{\rm d}m + \int_{m_{\rm s2}}^{m_{\max}} m\xi_3(m)\,{\rm d}m + m_{\max}, \tag{3.8}$$

where $m_{\rm s1}$ and $m_{\rm s2}$ denote the break points of the IMF. After substituting the corresponding power-law expression, we obtain analytical forms for each segment. For example, for the segment where $\alpha \neq 2$, the form of the integral is

$$\int m \cdot k m^{-\alpha}\,{\rm d}m = k \int m^{1-\alpha}\,{\rm d}m = k\left[\frac{m^{2-\alpha}}{2-\alpha}\right]. \tag{3.9}$$

Based on this, we then obtain the value of $m_{\max}$ and the IMF normalization constant. Once $m_{\max}$ and the normalization constant have been determined, the remaining stellar masses are generated iteratively. For a power-law segment of the IMF, the next integration boundary is obtained from

$$m_{i+1} = \left(m_i^{1-\alpha} - \frac{1-\alpha}{k}\right)^{\frac{1}{1-\alpha}}, \tag{3.10}$$

which follows directly from the condition that each interval of the IMF contains exactly one star. In this way, the stellar population is constructed sequentially from the most massive star towards lower masses.

It is important to note that the algorithm also takes into account the broader galactic context and incorporates additional factors that influence the final form of the IMF. These parameters include the influence of metallicity and environmental density, which is incorporated into the expression for α_3, whose value changes due to this effect for stars more massive than $1\,M_\odot$. The algorithm also incorporates the concept of a self-regulating process, which manifests itself in the dependence of $m_{\max}$ on $M_{\rm cl}$. Although the occurrence

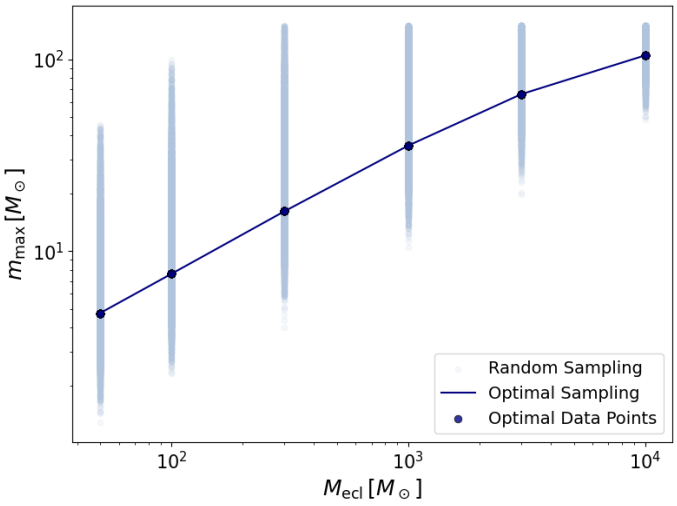


Figure 3.2: The relation between $m_{\max}$ and $M_{\rm cl}$. The light gray points represent random sampling, where we see a wide variation across the values of $m_{\max}$. The dark blue points represent optimal sampling values. In contrast to random sampling, optimal sampling defines a unique and deterministic value of $m_{\max}$ for each total mass $M_{\rm cl}$.

of stars with masses greater than $1\,M_\odot$ is not unusual in our calculations, we do not use the mentioned effects of metallicity and environmental density in the program and fix the coefficient α_3 to a constant value of 2.3. The concept of a self-regulating process is, of course, included here; without it, the optimal sampling calculation would not even make sense.

3.2 Properties of Optimal Sampling

As mentioned earlier, one of the key features of optimal sampling is the absence of any noise. For each value of $M_{\rm cl}$, there is exactly one specific star cluster. We selected six values of total $M_{\rm cl}$ (we chose the same values as in random sampling) and plotted the relation between $m_{\max}$ and $M_{\rm cl}$, see Figure 3.2. We see only six points, each corresponding to a single cluster size. In contrast to random sampling, there is no scatter here. Random sampling generates a wide range of values for $m_{\max}$ for the same $M_{\rm cl}$, introducing a relatively high degree of uncertainty into the model. Optimal sampling eliminates this variability, providing a much simpler way to study star clusters and galaxies more effectively and efficiently, without the need to simulate thousands of values to obtain an average value.

In random sampling, we work with the probability of a star of a certain mass occurring; optimal sampling, on the other hand, applies a strictly deterministic approach, a star either occurs or does not occur. The probability of a star of a certain mass occurring jumps abruptly from zero to one hundred percent. In Table 2.2, we have listed the probabilities of $m_{\max} \geq 20\,M_\odot$ for various $M_{\rm cl}$. Figure 3.3 then shows the theoretical minimum critical value for the observed $m_{\max}$. While the gray curve representing random sampling increases uncertainly and the probability of occurrence of $m_{\max}$ gradually increases, the navy curve sharply defines the threshold. The existence of such a massive star in systems below this

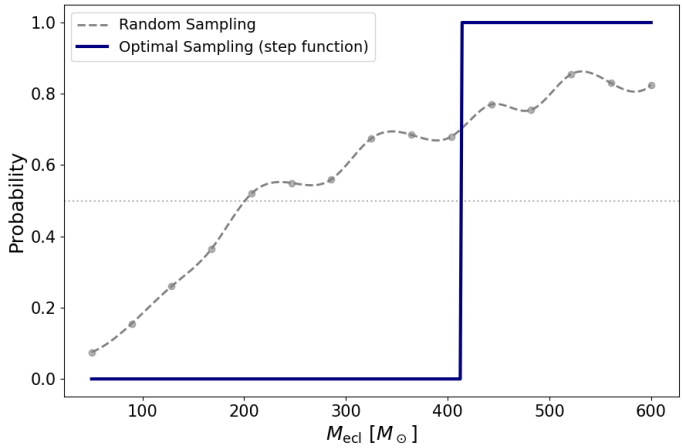


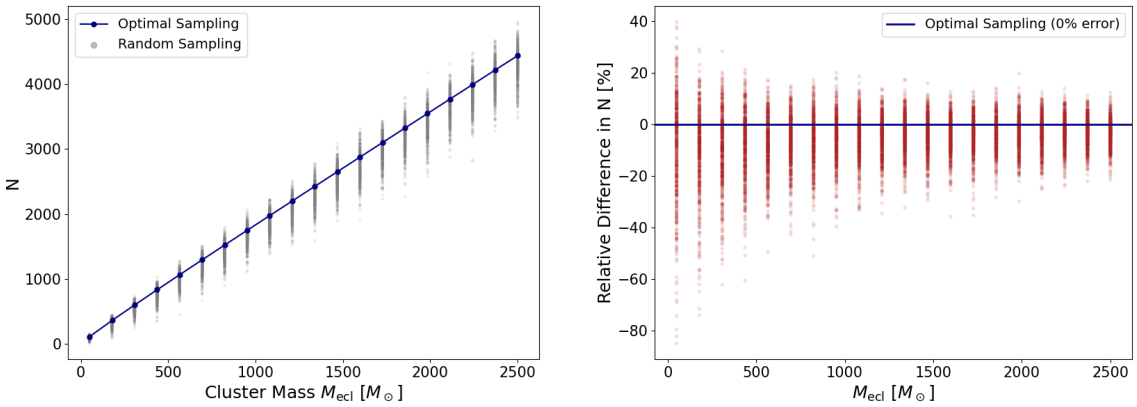
Figure 3.3: Comparison of the probability of finding a massive star ($m_{\rm max} \geq 20\,M_\odot$) as a function of the total mass of the star cluster $M_{\rm cl}$. The navy blue curve (optimal sampling) represents a sharp threshold for the occurrence of this star with a mass of $M_{\rm cl} = 412.59 M_\odot$. The gray curve (random sampling) shows a gradual increase in probability as $M_{\rm cl}$ increases, with the probability of $m_{\rm max} \geq 20\,M_\odot$ being approximately 73.7% for $M_{\rm cl} = 412.59\,M_\odot$.

threshold is ruled out in a deterministic model, which stands in sharp contrast to the results of random sampling, where even in low-mass systems there is a nonzero (even if low) probability of its formation.

According to optimal sampling, a massive star will never form in such a small star cluster, which would have a significant impact at higher levels. If massive stars do not form in low-mass star clusters, then in galaxies that form only small clusters (low SFR), the total IGIMF will be much steeper than in galaxies with high star formation rates. This then influences the chemical evolution (fewer supernovae, fewer heavy elements).

If we examine the variability/non-variability of individual approaches in more detail, it is necessary to point out how these approaches handle the total number of stars in the star cluster, N. If we work with a stochastic process under the condition of total mass $M_{\rm cl}$, N is random and variable; massive stars may or may not form in the cloud, and thus the number N generates noise of varying values. Optimal sampling eliminates this variability and replaces it with a single solution for each $M_{\rm cl}$, which provides stable and easily reproducible results. At the same time, however, it represents only an idealized limiting state, which the real universe is unlikely to achieve. We can say many beautiful things about the universe and describe it with various words, because that is exactly what it is. The star-formation process occurring within it is influenced by turbulence, magnetic fields, and local density fluctuations, and is therefore, in essence, more random than optimal. There are also many deviations from the idealized theoretical model, which, in its, shall we say, monotony, cannot be captured. There are cases that do not align with the relation $m_{\rm max} - M_{\rm cl}$ and are inaccessible to optimal sampling. Random sampling is more flexible in this regard; for it, these deviations are entirely natural.

We present the dependence of the number of stars on the total mass, illustrated in Figure 3.4a. Here we see a monotonic, linearly increasing deterministic curve generated by optimal sampling, and for each value of $M_{\rm cl}$, the number of scattered points modeled by random



(a) Absolute dispersion of N

(b) Relative dispersion (%)

Figure 3.4: Comparison of the number of stars (N) in stellar clusters as a function of cluster mass (M_{cl}). (a) The blue line represents modeling using optimal sampling, where for each M_{cl} we have a uniquely determined N. The gray points represent random sampling, where we performed 500 independent realizations for each value of M_{cl}. The spread of the gray points illustrates the influence of randomness in star formation, with this absolute variance increasing as M_{cl} increases. (b) Relative difference from the Optimal Sampling model, illustrating statistical convergence in accordance with the law of large numbers. The blue line represents the zero level, the ideal case resulting from optimal sampling, while the red points represent the dispersion resulting from random sampling.

sampling. The absolute deviation increases as M_{cl} increases, while the relative deviation decreases, which is consistent with the law of large numbers. The average of many random sampling runs thus approaches the expected value of the given IMF. Although the law of large numbers allows us to assume that, for massive stellar populations, the average across random realizations will approach the values obtained by optimal sampling, individual realizations still retain a stochastic nature and may exhibit fluctuations, particularly for quantities sensitive to rare massive stars. For low masses, the relative difference between the two approaches is greatest. This variability is a direct consequence of the stochastic nature of the IMF; in low-mass clusters, the random selection of even a single massive star can represent a substantial fraction of the available mass reservoir, which can lead to a much lower total number of stars compared to the deterministic value predicted by optimal sampling.

The question with which we might conclude this chapter is how significantly the two approaches differ in the overall shape of the stellar mass distribution. Can random sampling produce a realization that is comparably close to the reference IMF distribution as optimal sampling, or perhaps even closer? And if so, with what probability?

For comparison, we used the previously defined mass bins listed in Table 2.3. Each star cluster was described by a vector of the number fractions of stars in these bins, and the similarity between the two distributions was determined using the Euclidean distance between these vectors.

We worked with our dataset containing 10 000 random realizations for six different masses, and compared each of these 10 000 realizations with the corresponding optimally

Table 3.1: Comparison of the Euclidean distance of optimal sampling from the reference distribution with random sampling. The last column shows the empirical probability that a random realization reaches the same or a smaller distance than optimal sampling. Although the absolute values of the Euclidean distances are small, keep in mind that we are comparing normalized count fractions, so the relative difference between the two methods is significant. The median distance for random sampling is approximately 5–14 times greater than the corresponding distance for optimal sampling.

$M_{\rm cl}$ [$M_\odot$]	$d_{\rm opt}$	Median $d_{\rm rand}$	$P(d_{\rm rand} \leq d_{\rm opt})$ [%]
50	0.0127	0.0750	0.07
100	0.0079	0.0549	< 0.01
300	0.0033	0.0328	< 0.01
1000	0.0013	0.0182	< 0.01
3000	0.0021	0.0105	0.13
10000	0.0008	0.0058	0.01

sampled star cluster. For each star cluster, we first determined the distance of the optimally sampled star cluster from the reference distribution and designated it as the threshold distance. We then calculated how many random realizations achieved the same or a smaller distance from the same reference distribution, thereby obtaining an empirical estimate of the probability.

This is perhaps not surprising; such cases were either rare or absent within the finite simulation sample (our dataset could indeed be larger; in these cases, we can only say that their frequency is lower than the resolution limit determined by the size of the simulation dataset used). We present Table 3.1 here, which contains the obtained values.

In the set of random realizations for each cluster mass, only 0 to 0.13% of the realizations reached a distance from the reference distribution that was equal to or less than that of the corresponding optimally sampled cluster. The random realizations thus differ significantly from the smooth distribution in our dataset. If real star clusters systematically exhibited a distribution as smooth as that of optimal sampling, such star clusters would be very rare within purely stochastic random sampling, which could aid in comparison with real data.

Figure 3.5 illustrates the distance distribution between random sampling and optimal sampling from the reference distribution. It shows how far random sampling realizations generally deviate from the reference values, while optimal sampling lies significantly closer to the reference distribution (as expected). Although the distances of random sampling gradually decrease as the mass of the star cluster increases, the probability that random sampling will produce a realization as close as that of optimal sampling remains very low and does not systematically increase with mass.

In the limit of very massive star clusters, we expect random and optimal sampling to gradually converge, as relative stochastic fluctuations decrease with increasing star number. However, this limiting case does not address the situation of low-mass star clusters, for which the differences between the two approaches are most pronounced and debatable. These star clusters are, however, crucial within the IGIMF formalism, because in galaxies with low star formation rates, low-mass star clusters can constitute a significant portion of the total stellar population. The question of which sampling better describes their actual

mass distribution thus remains open.

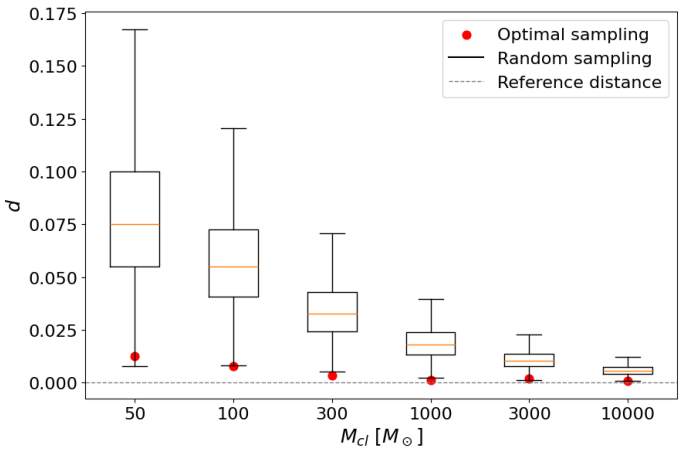


Figure 3.5: Distribution of Euclidean distances d between binned mass distributions of stars and the reference IMF. The boxplots show distances from 10 000 random sampling runs for each cluster mass $M_{\rm cl}$. The red dots represent the distances of the corresponding optimally sampled star clusters, and the dashed horizontal line represents the ideal reference value d=0. A gradual convergence of random and optimal sampling with increasing star cluster mass is evident, but so is the wide dispersion of individual random realizations.

Chapter 4

The Signature of Sampling

In the previous chapters, we examined two approaches to sampling stars and their properties. In this chapter, we will focus on whether it is possible to deduce, based on the data and individual characteristics, which sampling method was used to generate the stellar population. We will use machine learning methods for the classification itself, but first, we need to examine in more detail one key phenomenon outlined in Chapter 2. The fixed constraint on total mass has a significant impact on the resulting sample, and the distribution is not as random as it may seem. We will analyze this deviation in greater depth and examine how it can influence the internal structure of the data under analysis.

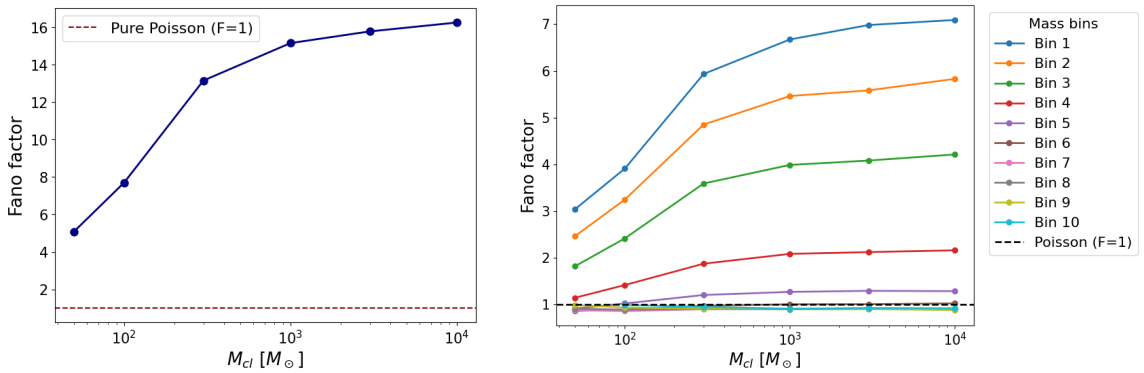
4.1 The Poissonian Behaviour

If counts in mass bins were generated by an independent Poisson process, we could view the distribution of stars as a Poisson distribution. This distribution is essentially quite simple and easy to apply, so using it would make things much easier in practice. The Poisson distribution is characterized by the equality between the mean μ and the variance σ^2, which reduces the entire problem to a single-parameter problem. This equality can be quantified using the so-called *Fano factor*, defined as

$$F = \frac{\sigma^2}{\mu}. \tag{4.1}$$

In the case of a purely independent random process, the Fano factor takes on the exact value of $F = 1$. However, as shown in Chapter 2, the restriction $M_{\rm cl}$ violates the assumption of independence, and the sampling algorithm itself thus enforces mutual dependence.

We therefore analyzed the total number of stars for the relevant values $M_{\rm cl}$. We calculated the average number of stars, the variance and the corresponding Fano factor in order to compare the behavior of N with the Poisson reference case. Table A.1 is provided in Appendix A and the values are illustrated in Figure 4.1a. The Fano factor increases with the size of the star cluster. Although for small star clusters ($M_{\rm cl} = 50\,M_\odot$) the Fano factor is of the order of unity, for massive star clusters ($M_{\rm cl} = 10000\,M_\odot$) it exceeds 16. It is significantly higher than the Poisson value $F = 1$ in all cases, which clearly shows that the total number of stars N at fixed $M_{\rm cl}$ is not a Poisson variable. This deviation is expected, since the fixed total mass couples the sampled stellar masses to the final value



(a) Fano factor for total number of stars

(b) Fano factor across mass bins

Figure 4.1: The dependence of the Fano factor on the target mass $M_{\rm cl}$. Subfigure (a) shows the Fano factor for the total number of stars. As the mass increases, a growing deviation between the calculated Fano factor and the Poisson factor ($F = 1$) becomes apparent. Subfigure (b) breaks down the dependence into individual mass bins. The excessive scatter is primarily caused by the lowest-mass stars, which have significantly higher Fano factor values (F = 4–7) in bins 1–4.

of N. This leads to compensating changes in the low-mass bins, which help maintain the fixed total-mass constraint when massive stars are present. The Fano factor reflects the fact that the absolute variance of N increases with cluster mass, while the relative scatter may still decrease. Thus, we used the Poisson distribution here mainly as a reference case against which the effect of the fixed-$M_{\rm cl}$ constraint can be compared.

Similar conclusions can be observed when we sum stars across mass bins, the range of which was shown in Table 2.3. If the individual bins followed a Poisson distribution, then their sum would also follow a Poisson distribution. However, as is evident from the previous calculations and also from the graph in Figure 4.1b, most bins deviate significantly from the Poisson distribution. Values less than one indicate suppressed dispersion (bins 7–10), while values greater than one indicate dispersion much higher than would correspond to the Poisson model (bins 1–4). The algorithm uses low-mass stars as "fillers" to precisely achieve the mass limit, while the formation of massive stars is strictly limited. The exact values are listed in Table A.2 and included in Appendix A.

In the studies by Cerviño et al. (2002) [17] and Cerviño and Valls-Gabaud (2003) [18], the authors examined the Poisson distribution and its application to small star clusters. They point out that in this case star clusters must be modeled on the basis of a fixed number of stars, and the total mass $M_{\rm cl}$ should be treated as a random variable. According to them, scaling using total mass is not appropriate for small systems because it does not account for the discrete nature of the population. It is not possible to fix both at the same time, and to maintain the Poisson model for small star clusters, it is necessary to model them using a fixed number N.

Our results support these theoretical conclusions and highlight the significant deviation from the ideal distribution that occurs when this assumption is violated. In astrophysical practice, it is often necessary to proceed in the opposite direction and use a fixed target mass of the star cluster as an input parameter. The total number N then becomes a random

variable, and the algorithm necessarily enforces a mutual dependence between the mass bins, resulting in strong correlations and extreme dispersion. It should be noted that these deviations are "actively encouraged" by our algorithm, which, in an effort to approximate the desired $M_{\rm cl}$, selects stars and thus alters the statistical nature of the problem. We therefore view the Poisson model here primarily as a reference model, which makes it easy to demonstrate the deviation caused by the sampling design.

4.2 Correlation

Before we get into the ML itself, let us take one more detour. In the previous section, we pointed out that individual bins influence each other. Now let us see just how strong this influence is. For a simple visualization, we present the correlation matrix for two cluster sizes, $M_{\rm cl} = 50\,M_\odot$ and $M_{\rm cl} = 10000\,M_\odot$ (Figure 4.2). The patterns are least and most pronounced for these selected $M_{\rm cl}$; with a larger dataset, individual relationships become clearer.

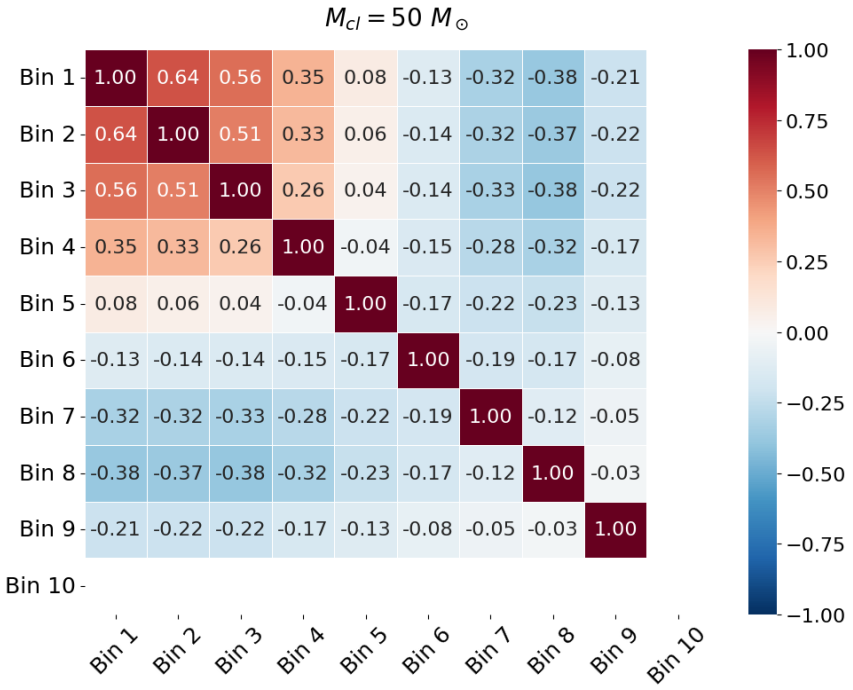
Nevertheless, both correlation matrices agree on the visible correlation pattern: low-mass bins form a "group" that, through their number of stars, compensates for the stochastic fluctuations of massive stars, so that an exact limit is achieved, which only confirms the previous analysis using the Fano factor. The negative correlation between the most massive and low-mass bins is therefore no longer surprising, but it certainly broadens our perspective.

Correlation matrices thus provide information about linear relationships between individual bins and show that deviations from Poissonian behavior are not merely independent effects of individual bins but are organized into a specific correlation pattern. This pattern is also relevant for subsequent machine learning analysis, as classification algorithms can utilize not only the values of individual characteristics, but also their mutual relationships.

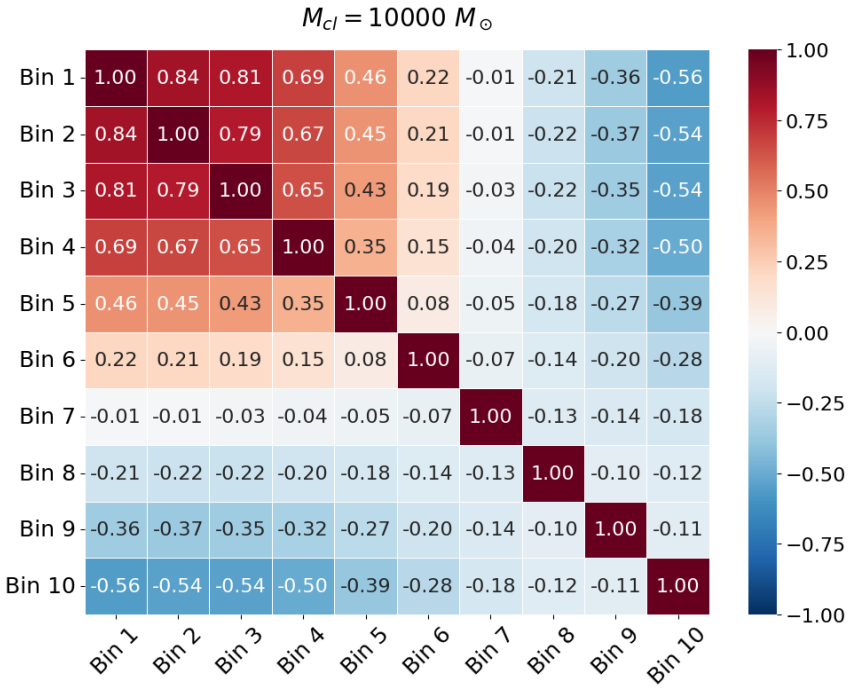
4.3 Machine Learning - Saving the Best for Last

After a few minor but interesting digressions, we come to the promised use of machine learning methods. As we have previously established, the limitation on the total mass $M_{\rm cl}$ has a significant impact not only on the variance of individual variables but also on the relationships between mass bins. We would like to further ask whether it is possible to determine, based on selected statistics, which sampling method was used to generate the stellar population, since information about the sampling method is likely not contained in a single parameter but in a combination of them. If there is a specific sampling mechanism governing the filling of the mass reservoir, it should leave a trace in the data that simple models cannot capture. We use the models we have selected for classification; at the same time, however, these models can provide us with indicative information about which characteristics contribute most to this distinguishability.

Let us not forget one important thing: we are using synthetic data generated for the purposes of this work. We are in no way working with real observations, and therefore we cannot answer the question of whether real observations can identify the physical sampling mechanism. If we were to apply this to real data, we would have to account for a number of



(a)



(b)

Figure 4.2: Correlation matrix of bins for masses of (a) 50 and (b) $10000\,M_\odot$. In both cases, a strong positive correlation is observed (up to 0.64 for $M_{\rm cl} = 50\,M_\odot$, up to 0.84 for $M_{\rm cl} = 10000\,M_\odot$) between bins 1–4, which behave collectively as a reservoir that fills or empties together depending on how much mass remains. Between these bins and the most massive bin 10 (in the case of $M_{\rm cl} = 50\,M_\odot$, bin 8), a significant negative correlation of up to -0.56 is then visible.

additional effects, such as unresolved binary stars, incompleteness, extinction, age spreads, stellar populations, contamination by field stars, measurement uncertainties... Our results thus demonstrate the distinguishability of both methods in an ideal simulated environment; we leave their applicability to real observations as a question for the future.

4.3.1 Datasets

For the analysis, we created two datasets:

1. *Balanced 1:1 dataset*

   For 10000 unique $M_{\rm cl}$ values in the range of $M_{\rm cl} = 50$–$10000\,M_{\odot}$, we generated one random sample and one optimal sample. Both classes are represented by the same number of realizations; however, the dataset does not fully capture the typical variability of random sampling.

2. *Unbalanced 1:20 dataset*

   This dataset covers the same mass range of $M_{\rm cl} = 50$–$10000\,M_{\odot}$, but for each of the 1000 selected values $M_{\rm cl}$, it contains one optimal realization and 20 independent realizations of random sampling. This corresponds to a 1:20 class ratio between optimal and random sampling. This unbalanced dataset better captures the stochastic variability of the random-sampling approach.

Each dataset was described by a set of features, but we used only some of them for modeling. Thus, we did not work directly with a list of individual stellar masses but rather with aggregated parameters describing the distribution of stellar masses within the population. We omitted variables that explicitly or implicitly depended on the size of the star cluster (number of stars, most massive star, or similar). These quantities could significantly simplify the classification; in such a case, the model could distinguish between individual methods based on the size of the system rather than on the actual shape of the mass distribution. We therefore chose primarily non-dimensional quantities and relative ratios, allowing the model to capture the relative structure of the distribution and compare star clusters across the entire $M_{\rm cl}$ range.

The 20 selected variables can generally be divided into several categories, a precise description of each is provided in Table B.1 in Appendix B.

1. The first group consists of number fractions in logarithmic mass bins, describing the relative occupation of the realized IMF.

2. The second group consists of cumulative mass fractions above selected thresholds (0.5, 1, 2, 4 and $8\,M_{\odot}$).

3. The third group comprises parameters describing the upper end of the IMF, specifically the dominance of the most massive stars. These include, for example, the mass fraction contained in the three or five most massive stars and the mass ratio of the two most massive stars.

4. The fourth group consists of a selected quantile ratio, `q75_over_q25`, which provides information about the relative width of the mass distribution.

5. The final feature is the luminosity ratio, defined as the fraction of the total luminosity contributed by the brightest star.

The target variable was the sampling label, which distinguished between random selection and optimal selection. The sampling label was used as the target variable in the supervised classification task, but was not included among the input features.

4.3.2 Classification Models

In this subsection, we will introduce the reader to the models we have chosen to use for our analysis. Some have proven to be more useful than others, so we will discuss them in greater detail later on, however, each model has contributed at least a small piece to the overall picture.

Dummy Classifier

First, we present a simple dummy classifier, which cannot really be considered a full-fledged model. This model does not use any input features and works only with the prior class distribution in the training data. For each object, it therefore returns the same class probabilities corresponding to their frequency in the training set. Thus, it served only as a baseline performance level, which we used to compare other models. If any of the more complex models failed to reach the baseline performance threshold, it would not have sufficient explanatory value for us.

Logistic Regression

The logistic regression model is one of the most widely used models for classification tasks, particularly in our case of binary classification. Unlike linear regression, it does not directly predict a continuous value, but rather the probability that an object belongs to a given class. This probability is obtained using the logistic function, also known as the sigmoid function

$$P(y=1|x) = \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}}, \tag{4.2}$$

which transforms a linear combination of input features into a value in the range from 0 to 1. Our model tests whether populations are approximately linearly separable in the space of selected characteristics. The advantage lies both in the simplicity of the model and its interpretability, as the magnitude and sign of the coefficient can be used to roughly determine which parameters contribute most to the classification. However, if some input features are mutually correlated, the interpretation of individual coefficients becomes less straightforward, and the results should therefore be interpreted with caution.

Random Forest

Another model is the random forest, which combines a set of decision trees. A single decision tree progressively splits the data based on the values of individual features so that the terminal branches contain the purest possible groups of a single class. A single

decision tree can be sensitive to the data and easily overfit; the random forest mitigates this overfitting by using a large number of decision trees trained on different subsets of the data. The final classification is then determined by a vote of the individual trees.

Unlike logistic regression, it is no longer limited to linear separation; it can thus capture more complex internal structures and interactions between input features, and also handle outliers more effectively. However, just as with logistic regression, it is also possible to roughly determine the importance of individual features here as well. The value of each feature indicates which variables the model uses the most when partitioning the data.

In our case, the depth of the trees was limited to reduce the risk of overfitting and to ensure the model retains a better ability to generalize to data outside the training sample. For an unbalanced dataset, we also used class balancing, where the model takes into account the different class distributions during training and does not automatically favor the more numerous class.

Support Vector Machine, SVM

The second-to-last model, and more of a supplementary one, is the support vector machine, SVM, which seeks a decision boundary that best separates the individual classes in the space of input features. In the simple linear case, this is a hyperplane that maximizes the distance between the two classes. The points closest to this boundary, the so-called support vectors, determine its final position. The model can also be applied to more complex, nonlinearly separable data by introducing various forms (including nonlinear) of kernel functions. In our work, we used the RBF (Radial Basis Function) kernel, also known as the Gaussian kernel. The RBF kernel measures the similarity between data points based on their Euclidean distance in the input space.

We used SVM primarily as a check on the discriminative power of the Random Forest.

K-Nearest Neighbors, KNN

The last model, which, like the SVM, served primarily to evaluate the classification performance of the Random Forest, is KNN, or k-nearest neighbors. Unlike the others, this model operates on a slightly different principle. It does not seek an explicit decision boundary, nor does it learn a set of coefficients, but rather classifies an object based on the classes of its nearest neighbors in the space of input features. In other words, a new object is classified based on which known objects it is closest to. The algorithm is thus based on distance and its use allows us to verify whether populations generated by the same sampling method form locally similar groups in the space of selected features.

KNN is sensitive to the degree of dispersion and overlap between classes. Based on the Fano factor from the previous section, which turned out to be quite high, this becomes important, especially in regimes where objects generated by different sampling methods overlap more strongly in the parameter space. In such cases, the nearest neighbors of a given object may not necessarily belong to the same class, which can reduce the performance of the model.

We selected a total of three models sensitive to the range of input variables: logistic regression, SVM, and KNN. Therefore, we began by standardizing the data. We rescaled the individual features so that they had a mean of zero and a standard deviation of one.

4.3.3 Validation Strategy

Before we get to the final results, let us summarize our testing and evaluation methodology. In general, we performed 4×3 tests for each dataset. By this, we mean that we conducted tests on four different sets of input features, for three different configurations of the training and test sets.

First, we worked with all the mentioned features; that is, we always used the full feature set for analysis. Next, we removed a specific group of features each time and observed how their absence affected the model's ability to distinguish between random and optimal sampling. It is called ablation testing:

- We removed the luminosity ratio feature to observe how much the results are driven by luminosity information,
- we removed information about the most massive stars to determine how much information is carried in the upper tail of the IMF,
- we left only the relative occupation of 10 logarithmic IMF bins to determine whether the sampling signature is contained in the binned shape of the realized IMF alone or only in specific derived features.

We split the data into a test and training set in three different ways. First, we performed a group random split in an 80:20 train-to-test ratio. The groups were defined based on the value $M_{\rm cl}$, so realizations corresponding to the same total mass were always assigned to the same part of the dataset. We performed this group random split a total of 50 times to mitigate the influence of a specific random division of the data (repeated group validation). We then tested two extrapolation methods: in one case, we trained on less massive star clusters and tested on more massive ones, and in the other case, we swapped these sets, meaning the model was trained on more massive star clusters and tested on less massive ones. The cutoff was 1000 $M_\odot$.

We compared the models using two metrics: ROC AUC and PR AUC. ROC AUC measures how well the model separates the two classes across different decision thresholds. The ROC curve itself describes the relationship between the true positive rate (TPR) and the false positive rate (FPR). Ideally, ROC AUC reaches a value of 1.0, which corresponds to perfect class separation. If a model has an ROC AUC of 0.5, it is as if it were classifying the data completely at random.

The PR AUC (Area Under the Precision-Recall Curve) expresses the trade-off between precision and recall. This metric focuses primarily on correctly identifying the positive class and is therefore particularly suitable for imbalanced datasets, and thus, in our case as well. In our implementation, optimal sampling is treated as the positive class 1, while random sampling is treated as the negative class 0.

For the repeated group random split, we calculated the average value of the metric and its standard deviation in all 50 repetitions for each model. These values gave us a better idea not only of the performance of the model but also of its stability with respect to the specific data distribution.

For extrapolation tests, we estimated the uncertainty of the resulting metrics using bootstrapping of the test set. This involves a form of sampling with replacement, which

means that the test set is resampled and the same data point may appear multiple times within it. The model is trained only once; we bootstrap only the test set and recalculate the metric value for each resampled selection. We then obtain an estimate of the uncertainty of the result from the variance of these values.

However, these metrics were not the only way we compared the models. We were also interested in their stability across repeated splits, the difference between testing in the full $M_{\rm cl}$ range and extrapolation tests, as well as the sensitivity of the results to a specific set of features. If a model achieved high performance only when using all features, but its performance dropped significantly after removing any group of features, this would mean that the classification is heavily dependent on that specific information. Conversely, if the model maintained good performance even after restricting the input features, this would suggest that the sampling signature is embedded more generally in the data structure.

More than the test across the entire range of $M_{\rm cl}$, we were interested in extrapolation tests. These answer the question of whether the learned difference between random and optimal sampling carries over to different mass ranges or whether the classification is tied only to a specific range of cluster masses.

4.3.4 Results and Interpretation

As we mentioned, we conducted A LOT of tests. We will list only the most important ones here; the remaining results and tables can be found in Appendix C. Here, we will primarily discuss models trained on a balanced dataset, since both classes are represented by the same number of instances, making the results easier to interpret. The unbalanced dataset then serves as a test of the robustness of these findings in cases where the stochasticity of random sampling is better captured.

The Full Feature Set

First, we analyze all five classification methods on the full feature set. Table 4.1 shows the performance of each method, and Table 4.2 shows the top five classification coefficients for logistic regression and random forest. To summarize, we highlight three main conclusions: all models outperform the baseline (which is good news); logistic regression proved to be the weakest, while RF, SVM, and KNN are nearly perfect, meaning that the samples are well distinguishable from each other based on the selected features.

Although logistic regression shows the weakest results, since they are still significantly higher than those of the dummy classifier, part of the sampling signature can be captured even using a simple linear model; however, a significant portion of the information is likely contained in the interrelationships that can be described by more complex models such as random forest, SVM, and KNN. The remaining three methods perform similarly, which suggests that the result is not specific to a single model type.

We might note that the performance of these methods is nearly equal to one, which usually occurs only in idealized or very well-separated cases. In our case, this is closely related to the deterministic nature of optimal sampling. For a given value of $M_{\rm cl}$, optimal sampling produces a unique stellar population and therefore a unique set of derived features. Across the full $M_{\rm cl}$ range, these points form a smooth and highly regular structure in the feature space. Random sampling, on the other hand, produces stochastic realizations

Model	ROC AUC	PR AUC
Dummy Classifier	0.4982	0.4991
Logistic Regression	0.8386 ± 0.0062	0.7365 ± 0.0099
Random Forest	0.9996 ± 0.0001	0.9996 ± 0.0002
SVM	0.9996 ± 0.0002	0.9994 ± 0.0003
KNN	0.9959 ± 0.0012	0.9919 ± 0.0023

Table 4.1: Model performance for the full feature set using repeated group validation. The reported values represent the mean and standard deviation over 50 repeated splits. The dummy classifier represents the baseline performance and is therefore reported without uncertainty.

Logistic Regression		Random Forest	
Feature	Coefficient	Feature	Importance
`m1_over_m2`	-7.594	`m1_over_m2`	0.215
`top3_mass_fraction`	-4.932	`bin_1_fraction`	0.108
`luminosity_ratio`	2.024	`bin_3_fraction`	0.096
`bin_10_fraction`	-2.016	`mass_fraction_above_0p5`	0.087
`top5_mass_fraction`	1.976	`mass_fraction_above_2`	0.084

Table 4.2: The five most significant variables for logistic regression and the random forest method for the full feature set. The values were obtained after training the models on the full dataset. Given the correlations among the input characteristics, these results must be interpreted with caution.

that scatter around this deterministic relation. The models are therefore not identifying determinism directly, but rather the statistical consequences of determinism: low scatter, regular feature relations, and systematic dependence on $M_{\rm cl}$.

If we look further at the most important features according to logistic regression and random forest, both models identify `m1_over_m2`, that is, the mass ratio of the two most massive stars, as one of the most significant parameters. From this we can infer that a significant amount of information is contained in the upper end of the IMF among the most massive stars, which also supports the importance of `bin_10_fraction`, `top3_mass_fraction`, `top5_mass_fraction`, and `luminosity_ratio` in logistic regression.

In addition to `m1_over_m2`, the random forest also uses number fractions in low-mass bins and the cumulative mass fraction above $0.5\,M_\odot$ and $2\,M_\odot$. The nonlinear model does not work solely with the upper end of the IMF, but also utilizes broader information about the realized IMF structure across the population, which is consistent with the previous correlation analysis (the low- and high-mass parts of the population are not entirely independent).

The following extrapolation represented a stricter form of model validation. It examined whether a model trained in one mass regime is capable of performing the same classification in another, i.e., a kind of transferability of the sampling signature. Based on Table 4.3, we see that all models exhibit a drop in performance compared to group validation. Thus, the selected mass range plays a significant role in training. Furthermore, we are in the "problematic" $M_{\rm cl}$ range, since, as we observed earlier, certain assumptions and regularities

Model	ROC AUC	PR AUC
Training: $M_{\rm cl} \leq 1000\,M_\odot$, *Test:* $M_{\rm cl} > 1000\,M_\odot$		
Dummy Classifier	0.5005 ± 0.0053	0.5005 ± 0.0067
Logistic Regression	0.7087 ± 0.0059	0.6823 ± 0.0090
Random Forest	0.6921 ± 0.0062	0.5856 ± 0.0079
SVM	0.7196 ± 0.0047	0.7623 ± 0.0052
KNN	0.6961 ± 0.0041	0.6925 ± 0.0062
Training: $M_{\rm cl} > 1000\,M_\odot$, *Test:* $M_{\rm cl} \leq 1000\,M_\odot$		
Dummy Classifier	0.5015 ± 0.0049	0.5011 ± 0.0056
Logistic Regression	0.7404 ± 0.0046	0.6834 ± 0.0065
Random Forest	0.9284 ± 0.0021	0.9356 ± 0.0020
SVM	0.5986 ± 0.0025	0.5988 ± 0.0044
KNN	0.7081 ± 0.0035	0.7086 ± 0.0042

Table 4.3: Extrapolation performance for the full feature set. The models were trained on one cluster-mass regime and tested on the other. The boundary between the two regimes was set to $M_{\rm cl} = 1000\,M_\odot$, and the uncertainties were estimated by bootstrapping the test set.

applicable to more massive clusters cease to hold for small clusters. Low-mass star clusters are more chaotic with greater fluctuations, making it more challenging to identify patterns that would be transferable to more massive star clusters.

The drop is particularly strong for the extrapolation from low-mass to high-mass clusters, where the random forest reaches only ROC AUC $= 0.6921$ and PR AUC $= 0.5856$. Logistic regression achieves similar performance in both directions, around a ROC AUC > 0.70. This indicates that the linear component of the sampling signature is relatively stable, but not strong enough for accurate classification outside the training mass range. SVM and KNN are sensitive to the geometry of the training data, which may explain why their performance on the test set drops significantly compared to group validation (SVM: $0.9996 \rightarrow 0.7196/0.5986$ and KNN: $0.9959 \rightarrow 0.6961/0.7081$). In summary, the previous, nearly perfect performance in cross-validation was largely due to the fact that both training and test sets operated within the same mass range.

If we take a closer look at the random forest, we notice that it performs better when extrapolating from large to small star clusters (ROC AUC 0.93) than in the opposite direction (ROC AUC 0.69). One possible interpretation is that, when trained on more massive star clusters, the model learns a more stable and less noisy sampling signature, which it is subsequently able to recognize even in the more variable environment of low-mass star clusters. Conversely, when trained on small star clusters, the model works with noisier data, so it may have difficulty capturing patterns that would be just as clearly transferable to the more stable environment of more massive star clusters.

No Luminosity Set

For this and subsequent ablation studies, we used only logistic regression and random forest. In this section, we are no longer primarily testing overall classification performance, but rather observing how model performance changes after removing selected groups of input

Test	Logistic Regression	Random Forest
Group validation	0.6303 ± 0.0068	0.9992 ± 0.0004
Small $\rightarrow$ Large	0.5684 ± 0.0059	0.7481 ± 0.0063
Large $\rightarrow$ Small	0.6016 ± 0.0059	0.9382 ± 0.0020

Table 4.4: Ablation test without the `luminosity_ratio` feature. The table shows PR AUC values for logistic regression and random forest. The decline compared to group validation is noticeable especially for logistic regression, while random forest remains robust, particularly in the extrapolation from high-mass to low-mass clusters.

Small $\rightarrow$ *Large*			
Logistic Regression		Random Forest	
Feature	Coef.	Feature	Imp.
`bin_10_fraction`	-3.948	`m1_over_m2`	0.211
`bin_9_fraction`	-3.418	`bin_3_fraction`	0.121
`m1_over_m2`	-1.690	`bin_1_fraction`	0.112
Large $\rightarrow$ *Small*			
Logistic Regression		Random Forest	
Feature	Coef.	Feature	Imp.
`bin_10_fraction`	-1.226	`bin_4_fraction`	0.122
`mass_fraction_above_8`	-0.864	`mass_fraction_above_4`	0.100
`m1_over_m2`	-0.785	`bin_3_fraction`	0.097

Table 4.5: Top three features in the extrapolation tests after removing the `luminosity_ratio` feature. Logistic regression values correspond to coefficients, while random forest values correspond to feature importances.

features. For comparison, we therefore use only PR AUC. Compared to ROC AUC, this metric is more sensitive to the quality of positive class classification and therefore better highlights any decline in performance. We also focus more on top features in extrapolation models because we are interested in which features the model uses when transferring between different mass ranges.

The `luminosity_ratio` proved to be a significant linear indicator for the full feature set (under ideal conditions), so we are interested in whether the sampling signature remains sufficiently robust even if we completely eliminate this feature.

After removing `luminosity_ratio`, the decline is indeed noticeable, particularly in the case of logistic regression, as is shown in Table 4.4. Its classification performance during extrapolation is very similar in both directions; moreover, in both cases, massive stars, or rather, the upper end of the IMF, play a significant role.

With random forest, the performance for group validation remains almost as high as for the full feature set. However, when it comes to extrapolating from low-mass star clusters to high-mass ones, performance also declines. In the opposite direction, the decline is nowhere near as pronounced; the nonlinear model is thus able to compensate for the missing luminosity information to some extent by combining other parameters, such as the ratio of the most massive stars, cumulative mass fractions, or the number fractions in individual mass bins.

Test	Logistic Regression	Random Forest
Group validation	0.6380 ± 0.0078	0.9976 ± 0.0008
Small $\rightarrow$ Large	0.5794 ± 0.0072	0.6108 ± 0.0062
Large $\rightarrow$ Small	0.5710 ± 0.0060	0.9515 ± 0.0016

Table 4.6: Ablation test after removing the features associated with the upper IMF extremes. For logistic regression, the performance in extrapolation tests decreases substantially and approaches the level of random guessing. For the random forest, the overall performance remains significantly higher, but the decline is particularly noticeable when extrapolating from low-mass clusters to high-mass clusters.

Small $\rightarrow$ *Large*			
Logistic Regression		Random Forest	
Feature	Coef.	Feature	Imp.
`bin_10_fraction`	-5.807	`bin_1_fraction`	0.166
`bin_9_fraction`	-3.207	`bin_2_fraction`	0.123
`mass_fraction_above_8`	-1.149	`mass_fraction_above_0p5`	0.107
Large $\rightarrow$ *Small*			
Logistic Regression		Random Forest	
Feature	Coef.	Feature	Imp.
`bin_10_fraction`	-0.917	`bin_4_fraction`	0.119
`mass_fraction_above_8`	-0.804	`mass_fraction_above_8`	0.115
`mass_fraction_above_1`	0.686	`bin_3_fraction`	0.105

Table 4.7: Top three features in the extrapolation tests after removing the features associated with the upper IMF extremes. Logistic regression values correspond to coefficients, while random forest values correspond to feature importances.

`Luminosity_ratio` thus contributes to the classification of the sampling method and carries some linear information, which is particularly evident in logistic regression; however, it is not the only one, and even without it, the random forest is able to maintain high classification performance.

No Upper-IMF Extremes Set

As was evident from Table 4.2, massive stars play a significant role in classification; therefore, it is not surprising that the classification performance of the models declines when parameters describing the upper end of the IMF are removed. As shown in Table 4.6, the performance of logistic regression for group validation reaches a value similar to that when the luminosity ratio is removed; however, in the case of extrapolation, it reaches a level almost comparable to random guessing. The coefficients listed in Table 4.7 show that even after removing extreme parameters, the model attempts to utilize information from more massive bins, particularly `bin_10_fraction`, `bin_9_fraction`, and `bin_8_fraction`; however, without information about luminosity and dominance of the most massive stars, these parameters do not carry sufficient linearly transferable information.

The Random Forest still maintains its high performance for group validation, just as it

does when extrapolating from more massive star clusters to less massive ones; however, like logistic regression, its performance drops substantially in the opposite direction of extrapolation. Its performance here is comparable to that of logistic regression. Unlike logistic regression, the random forest in this case relies primarily on the number fractions in the least massive bins, which can be interpreted as an attempt to use the broader binned shape of the realized IMF.

Removing the parameters describing the upper end of the IMF thus represented a loss of information for both models, and for logistic regression, this was truly significant, as it likely eliminates most linearly recognizable and transferable dependencies. The random forest is more robust, so it coped better with the loss of information; in any case, these parameters were also important for it, primarily when extrapolating from small star clusters to large ones.

The Bins-only Set

Finally, we were interested in how much information is contained solely in the number fractions of individual logarithmic mass bins. In other words, whether the difference between random and optimal sampling is already visible in the shape of the IMF itself.

Even in this limited feature set, the random forest maintains very high performance, both in group validation and in both extrapolation directions (Table 4.8). Compared to previous ablation sets, its performance is actually the most stable. This may suggest that the information about sampling is significantly embedded in the IMF shape; other derived features may have merely reproduced the information already present and did not necessarily improve the transferability of the model.

For logistic regression, performance is again substantially lower than that of the random forest; in extrapolation tests, it remains only moderately above random guessing in both directions. As with the previous non-extremes set, the most significant coefficients are primarily associated with the most massive bins, as is illustrated in Table 4.9. Even in this bins-only dataset, the linear model attempts to distinguish between sampling methods primarily using the upper end of the IMF.

The random forest behaves somewhat differently. When extrapolating from small to large, it relies mainly on the lightest bins, particularly `bin_1_fraction`, `bin_2_fraction`, and `bin_3_fraction`. In the opposite direction, the significant features shift more toward the medium-mass bins, mainly `bin_3_fraction`, `bin_4_fraction`, and high-mass bin `bin_10_fraction`. The nonlinear model does not rely on a single extreme parameter, but on a broader pattern in the relative occupation of the IMF bins; although a high-mass bin still appears in one extrapolation direction, the dominant information is carried by low- and intermediate-mass bins.

Based on the poor classification performance of logistic regression, we again conclude that the shape of the mass distribution alone does not carry sufficient information that would be easily captured by a linear model. However, the random forest performs surprisingly well, indicating that when nonlinear relationships and interactions between bins are taken into account, the shape of the IMF itself is sufficiently informative to distinguish between the two sampling methods.

Test	Logistic Regression	Random Forest
Group validation	0.6230 ± 0.0064	0.9969 ± 0.0009
Small $\rightarrow$ Large	0.5878 ± 0.0073	0.9277 ± 0.0032
Large $\rightarrow$ Small	0.5809 ± 0.0062	0.9723 ± 0.0012

Table 4.8: Ablation test using only number fractions in logarithmic IMF bins. The table shows PR AUC values for logistic regression and random forest.

Small $\rightarrow$ *Large*			
Logistic Regression		Random Forest	
Feature	Coef.	Feature	Imp.
`bin_10_fraction`	-4.980	`bin_1_fraction`	0.306
`bin_9_fraction`	-2.641	`bin_2_fraction`	0.247
`bin_8_fraction`	-0.629	`bin_3_fraction`	0.231
Large $\rightarrow$ *Small*			
Logistic Regression		Random Forest	
Feature	Coef.	Feature	Imp.
`bin_10_fraction`	-1.470	`bin_4_fraction`	0.207
`bin_9_fraction`	-0.104	`bin_3_fraction`	0.196
`bin_8_fraction`	-0.091	`bin_10_fraction`	0.180

Table 4.9: Top three features in the extrapolation tests using only number fractions in logarithmic IMF bins. Logistic regression values correspond to coefficients, while random forest values correspond to feature importances.

4.3.5 Final Summary

Let us summarize the overall evaluation of the results in a few paragraphs. Random and optimal sampling are clearly distinguishable from one another in the selected feature space and carry different linear and nonlinear information about their structure. This conclusion was to some extent expected, as the full feature dataset contained very strong indicators of the upper end of the IMF, such as `m1_over_m2`, `luminosity_ratio`, and the mass fractions of the most massive stars. More useful and interpretatively interesting, therefore, were the ablation tests and extrapolations, which show where the sampling signature is actually embedded and how well it is transferable across different mass regimes.

Extrapolation tests show that the sampling signature is not entirely universal across the entire M_{cl} range. This asymmetry is most clearly visible for the random forest, which generally performs better when trained on high-mass clusters and tested on low-mass clusters than in the opposite direction. Models trained in one mass regime did not achieve the same performance when applied to another regime, which could suggest that the difference between random and optimal sampling takes a slightly different form in low-mass and more massive star clusters. The more stable structure of more massive systems provides a clearer sampling signature, which is then better recognized in low-mass star clusters.

At the same time, it appears that information about the sampling method is not contained in just one variable. The `luminosity_ratio` and parameters describing the upper end of

the IMF do carry a significant portion of this information (especially in logistic regression), but removing them does not destroy the classification, as was shown with the random forest.

Let us highlight the conclusions from the bins-only case as the most interesting result. Even after removing strong derived indicators and retaining only number fractions in individual IMF bins, the model maintained high performance. The sampling signature is thus already present in the very shape of the realized IMF. What is surprising, however, is that in extrapolation tests, the random forest does not rely solely on the most massive bins, but also makes strong use of the low- to intermediate-mass portion of the distribution. These bins therefore likely carry indirect information about whether and how strongly the upper end of the IMF is populated, through the constraints it imposes on the remaining stellar population.

From a practical standpoint, this could mean that the sampling signature need not be sought solely through the most massive stars. Even in cases where the upper end of the IMF is not well known or is subject to high uncertainty, some of the information may be contained in the broader binned number distribution of stellar masses. At the same time, however, it is necessary to emphasize that these conclusions are based on controlled synthetic data. Real observations would introduce further complications, such as data incompleteness, errors in mass estimates, unresolved multiple systems, the dynamical evolution of the star cluster, or dependence on the age of the population.

Conclusions

As is customary at the end, we summarize the content of our work here. We were interested in two approaches to IMF sampling, random and optimal sampling, which can be understood as opposites in their interpretation of how the continuous IMF is realized in a specific stellar population, with the actual sampling process likely falling somewhere in between. Our goal was to compare these approaches and assess whether it is possible to find a signature in the data that would indicate how a given star cluster was sampled. We performed all analyses on synthetic data; that is, we did not attempt to directly identify the physical sampling mechanism in real observations. We worked with star clusters with masses ranging from 50 to 10000 $M_\odot$, as these differences are most pronounced in small star clusters.

We first addressed both approaches separately to familiarize readers with their basic characteristics. We began with the older and more widely used random sampling method, which employs the IMF as a probability density function from which stellar masses are selected at random. In the most basic interpretation, we work with random selection from a fixed number of stars N; for a physical interpretation, however, it is more meaningful to work with a fixed total mass $M_{\rm cl}$. However, a fixed mass introduces relationships in the data that then influence the statistical nature of the selection; in our case, this is further amplified by the chosen sampling algorithm. The final population is therefore no longer equivalent to an unrestricted random selection from the IMF. In particular, the random selection of a very massive star in small star clusters strongly influences the remaining available mass and thus the total number of stars N, leading to strong correlations between $m_{\rm max}$ and N. These correlations are also reflected in the Fano factor analysis. The total number of stars N at fixed $M_{\rm cl}$ does not behave as a Poisson variable, because the fixed total mass couples the selected stellar masses to the final value of N. The increasing Fano factor therefore indicates growing absolute variance in N, while the relative scatter can still decrease for more massive clusters.

Optimal sampling, on the other hand, represents a fully deterministic approach. It establishes a unique relation between $m_{\rm max}$ and $M_{\rm cl}$, and its goal is to reproduce the theoretical IMF as faithfully as possible. Unlike random sampling, it is not affected by Poisson noise; that is, for every value of $M_{\rm cl}$, there is one corresponding star cluster realization. Optimal sampling thus represents an idealized limiting prescription rather than a random realization of a stellar population.

The final chapter then focused on distinguishing between these two approaches using machine learning methods. The differences between the sampling prescriptions are not limited to a single property, such as the most massive star, the number of stars, and so on. Classification methods were able to distinguish very clearly in most cases how a given

star cluster was sampled. Nonlinear methods (random forest, SVM, KNN) were more successful; while logistic regression did capture some linearly transferable information, the performance of this method was significantly lower. The strongest information was contained in the high-mass end of the IMF, the mass ratio of the most massive stars, the luminosity ratio of the brightest star to the total cluster luminosity, or the mass fraction of massive stars. However, once we omitted this information about the extremes, it turned out that the sampling signature is also present in the broader shape of the realized IMF; the number distribution of stars in the low- and intermediate-mass bins emerged as significant factors. This suggests that the difference between random and optimal sampling is encoded in the internal structure as a whole.

These results therefore suggest that the sampling method can leave a detectable statistical imprint on the structure of a stellar population, at least within the controlled framework considered in this thesis. This result should, however, be interpreted with caution. Correlations between individual characteristics may have introduced additional information into the model that could have skewed the results, particularly in the case of logistic regression. For real stellar populations, we would then also have to take into account other factors, including unresolved binaries, incompleteness, extinction, age spreads, stellar evolution, field contamination, and measurement uncertainties.

These results are important not only for the modeling of individual star clusters, but also for the interpretation of more complex stellar populations. If different sampling prescriptions leave distinguishable signatures in the stellar mass distribution, then the way in which the IMF is populated may influence quantities that are commonly used to interpret unresolved or semi-resolved stellar systems. If we extend our considerations to semi-resolved stellar populations, the issue of IMF sampling becomes even more important. Since not all stars are observable individually, the brightest, often massive stars may be distinguished separately, while the rest of the population contributes only to the overall unresolved light. If the presence or absence of a few massive stars differs significantly between random and optimal sampling, this difference may also affect the interpretation of semi-resolved populations. However, since this study focused primarily on the IMF at the star cluster level, extending the analysis to semi-resolved stellar populations is a future application.

Appendix A

$M_{\rm cl}$ [$M_\odot$]	$\langle N \rangle$	Var(N)	Fano factor
50	101.56	516.65	5.09
100	189.93	1461.38	7.69
300	532.16	6994.44	13.14
1000	1722.91	26 085.76	15.14
3000	5139.10	81 018.92	15.77
10000	17 091.16	277 536.59	16.24

Table A.1: Table of Fano factor values for various $M_{\rm cl}$. We can see that as N increases, the dispersion increases significantly, and consequently so does the corresponding Fano factor.

	$M_{\rm cl}$ [$M_\odot$]					
Bin	50	100	300	1000	3000	10000
bin 1	3.046	3.913	5.935	6.674	6.983	7.091
bin 2	2.470	3.249	4.854	5.464	5.585	5.831
bin 3	1.826	2.419	3.593	3.990	4.084	4.214
bin 4	1.150	1.421	1.878	2.088	2.126	2.164
bin 5	0.918	1.027	1.208	1.275	1.297	1.291
bin 6	0.868	0.897	0.979	1.010	1.011	1.029
bin 7	0.881	0.864	0.905	0.906	0.906	0.912
bin 8	0.915	0.894	0.907	0.904	0.911	0.916
bin 9	0.992	0.939	0.917	0.913	0.906	0.886
bin 10	–	0.996	0.953	0.915	0.930	0.923

Table A.2: Fano factor values for the number of stars in individual mass bins and for various values of the target mass $M_{\rm cl}$. A dash indicates a case where the Fano factor could not be determined.

Appendix B

Category	Feature	Definition	Meaning
Binned IMF shape	`bin_1_fraction`–`bin_10_fraction`	$f_i = N_{\mathrm{bin},i}/N_{\mathrm{tot}}$	Relative occupation of logarithmic IMF bins
Upper IMF	`top3_mass_fraction`	$\sum_{j=1}^{3} m_j/M_{\mathrm{cl}}$	Three most massive stars
	`top5_mass_fraction`	$\sum_{j=1}^{5} m_j/M_{\mathrm{cl}}$	Five most massive stars
Mass thresholds	`mass_fraction_above_0p5`	$M(m>0.5)/M_{\mathrm{cl}}$	Mass above $0.5\,M_{\odot}$
	`mass_fraction_above_1`	$M(m>1)/M_{\mathrm{cl}}$	Mass above $1\,M_{\odot}$
	`mass_fraction_above_2`	$M(m>2)/M_{\mathrm{cl}}$	Mass above $2\,M_{\odot}$
	`mass_fraction_above_4`	$M(m>4)/M_{\mathrm{cl}}$	Mass above $4\,M_{\odot}$
	`mass_fraction_above_8`	$M(m>8)/M_{\mathrm{cl}}$	Mass above $8\,M_{\odot}$
Statistical ratios	`m1_over_m2`	m_1/m_2	Ratio of the two most massive stars
	`q75_over_q25`	$Q_{0.75}/Q_{0.25}$	Width of the mass distribution
Physical ratio	`luminosity_ratio`	$L_{\mathrm{max}}/L_{\mathrm{tot}}$	Fraction of total luminosity contributed by the brightest star

Table B.1: An overview of selected features used for classification via machine learning methods. For the final analysis, we used only non-dimensional variables and relative ratios. We did not include other absolute parameters so that the model would not be driven solely by the size of the system.

Appendix C

Model	ROC AUC	PR AUC
Repeated group validation, 50 splits		
Dummy Classifier	0.5119	0.0492
Logistic Regression	0.8320 ± 0.0047	0.1290 ± 0.0043
Random Forest	0.9993 ± 0.0004	0.9906 ± 0.0032
SVM	0.9981 ± 0.0005	0.9533 ± 0.0141
KNN	0.9987 ± 0.0004	0.9494 ± 0.0160
Training: $M_{\rm cl} \leq 1000\,M_\odot$, *Test:* $M_{\rm cl} > 1000\,M_\odot$		
Dummy Classifier	0.4995 ± 0.0051	0.0478 ± 0.0025
Logistic Regression	0.7096 ± 0.0118	0.1158 ± 0.0090
Random Forest	0.7767 ± 0.0047	0.1363 ± 0.0071
SVM	0.7229 ± 0.0136	0.4572 ± 0.0201
KNN	0.7007 ± 0.0108	0.2212 ± 0.0176
Training: $M_{\rm cl} > 1000\,M_\odot$, *Test:* $M_{\rm cl} \leq 1000\,M_\odot$		
Dummy Classifier	0.4998 ± 0.0046	0.0479 ± 0.0022
Logistic Regression	0.7384 ± 0.0076	0.0986 ± 0.0061
Random Forest	0.9173 ± 0.0053	0.6337 ± 0.0173
SVM	0.6013 ± 0.0084	0.2399 ± 0.0161
KNN	0.7032 ± 0.0109	0.4292 ± 0.0212

Table C.1: Classification performance for an unbalanced dataset with a 1:20 ratio using the full feature set. The first section shows repeated group validation over 50 splits. The second and third blocks show extrapolations between clusters with lower and higher masses, where we chose $M_{\rm cl} = 1000\,M_\odot$ as the threshold. The uncertainties in the extrapolation tests were estimated using bootstrapping of the test set.

For the sake of completeness, we also present the results of testing on a 1:20 unbalanced dataset. Here, we consider the case where all 20 features are available. The performance metrics for each test (validation and extrapolation) are shown in Table C.1. Optimal sampling is treated as the positive class; therefore, for the 1:20 unbalanced dataset, the dummy PR AUC is expected to be close to the positive-class prevalence, $1/21 \simeq 0.048$. As is evident at first glance, the ROC AUC values remain within a similar range as for the balanced dataset. However, a significant change occurs in the PR AUC, which is a much stricter metric in the case of imbalanced classes because it more strongly reflects the ability of the model to correctly capture the minority class.

In the case of group validation, the PR AUC values for methods that capture nonlinear information remain consistently high, close to the level of perfect classification; however,

the first significant change occurs with logistic regression. Its PR AUC is only slightly higher than that of the dummy classifier, suggesting that, in an unbalanced dataset, the linearly transferable information is much less useful for identifying the minority class.

However, as soon as we move to extrapolation, the performance of all methods drops substantially. When extrapolating from low-mass to high-mass clusters, the performance of the random forest, which otherwise achieved the best results in most previous tests, declines noticeably. Surprisingly, SVM performed best, though even here its performance is not particularly high. When the direction of extrapolation is reversed, that is, when training on higher-mass clusters and testing on lower-mass clusters, the ranking of the models changes again, and the random forest achieves the best result. Compared to the baseline of 0.0479, the value of 0.6337 is significantly higher, indicating that the model still captures meaningful information contained in the data. While the other models also perform above the baseline, their performance is significantly lower than that of the random forest.

The results show that the imbalance in the dataset has a significant impact on the interpretation of classification performance. The models are still capable of capturing information about the sampling method, but the task is significantly more challenging than in the balanced case. With a realistically unbalanced class distribution, classification is more sensitive not only to the metric used but also to the direction of extrapolation between different mass regimes.

References

[1] P. Kroupa, *The current astrophysical understanding of the initial mass function*. [Online video]. YouTube, 7 April 2022. Available from: https://www.youtube.com/watch?v=Sxtk2cIlU14&t=2457s [cit. 12. 5. 2026].

[2] P. Kroupa, *On the Variation of the Initial Mass Function*, Monthly Notices of the Royal Astronomical Society, vol. 322, no. 2, pp. 231–246, 2001. doi: 10.1046/j.1365-8711.2001.04022.x.

[3] E. E. Salpeter, *The Luminosity Function and Stellar Evolution*, The Astrophysical Journal, vol. 121, pp. 161–167, 1955. doi: 10.1086/145971.

[4] Y. Terzian, *Edwin Ernest Salpeter. 3 December 1924 – 26 November 2008*, Biographical Memoirs of Fellows of the Royal Society, vol. 56, pp. 391–399, 2010. doi: 10.1098/rsbm.2010.0005.

[5] P. Kroupa and T. Jerabkova, *The Salpeter IMF and its descendants*, Nature Astronomy, vol. 3, pp. 482–484, 2019. doi: 10.1038/s41550-019-0793-0.

[6] G. E. Miller and J. M. Scalo, *The Initial Mass Function and Stellar Birthrate in the Solar Neighborhood*, Astrophysical Journal Supplement Series, vol. 41, pp. 513–547, 1979. doi: 10.1086/190629.

[7] P. Kroupa, C. Weidner, J. Pflamm-Altenburg, et al., *The Stellar and Sub-Stellar Initial Mass Function of Simple and Composite Populations*, in *Planets, Stars and Stellar Systems*, vol. 5, Springer, Dordrecht, 2013, pp. 115–242. doi: 10.1007/978-94-007-5612-0_4.

[8] G. Chabrier, *Galactic Stellar and Substellar Initial Mass Function*, Publications of the Astronomical Society of the Pacific, vol. 115, pp. 763–795, 2003. doi: 10.1086/376392.

[9] C. Weidner and P. Kroupa, *The maximum stellar mass, star-cluster formation and composite stellar populations*, Monthly Notices of the Royal Astronomical Society, vol. 365, pp. 1333–1347, 2006. doi: 10.1111/j.1365-2966.2005.09824.x.

[10] M. R. Haas and P. Anders, *Variations in integrated galactic initial mass functions due to sampling method and cluster mass function*, Astronomy & Astrophysics, vol. 512, A79, 2010. doi: 10.1051/0004-6361/200912967.

[11] N. Bastian, K. R. Covey, and M. R. Meyer, *A Universal Stellar Initial Mass Function? A Critical Look at Variations*, Annual Review of Astronomy and Astrophysics, vol. 48, pp. 339–389, 2010. doi: 10.1146/annurev-astro-082708-101642.

[12] D. F. Figer, *An upper limit to the masses of stars*, Nature, vol. 434, pp. 192–194, 2005. doi: 10.1038/nature03293.

[13] P. Hennebelle and L. Grudić, *The Physical Origin of the Stellar Initial Mass Function*, Annual Review of Astronomy and Astrophysics, vol. 62, pp. 1–46, 2024. doi: 10.1146/annurev-astro-052622-031748.

[14] E. Applebaum, A. M. Brooks, T. R. Quinn, and C. R. Christensen, *A stochastically sampled IMF alters the stellar content of simulated dwarf galaxies*, Monthly Notices of the Royal Astronomical Society, vol. 492, pp. 8–21, 2020. doi: 10.1093/mnras/stz3331.

[15] C. Schulz, J. Pflamm-Altenburg, and P. Kroupa, *Mass distributions of star clusters for different star formation histories in a galaxy cluster environment*, Astronomy & Astrophysics, vol. 582, A93, 2015. doi: 10.1051/0004-6361/201526063.

[16] Z. Yan, T. Jerabkova, and P. Kroupa, *The optimally sampled galaxy-wide stellar initial mass function: observational tests and the publicly available GalIMF code*, Astronomy & Astrophysics, vol. 607, A126, 2017. doi: 10.1051/0004-6361/201730987.

[17] M. Cerviño, D. Valls-Gabaud, V. Luridiana, and J. M. Mas-Hesse, *Confidence levels of evolutionary synthesis models II. On sampling and Poissonian fluctuations*, Astronomy & Astrophysics, vol. 381, pp. 51–64, 2002. doi: 10.1051/0004-6361:20011266.

[18] M. Cerviño and D. Valls-Gabaud, *On biases in the predictions of stellar population synthesis models*, Monthly Notices of the Royal Astronomical Society, vol. 338, pp. 481–496, 2003. doi: 10.1046/j.1365-8711.2003.06068.x.