Ph.D Theses

Sample Size Problems in Extreme Value Statistics and the Luminosity Distribution of Galaxies

Katalin Ozogány

Supervisors: Dr. Zoltán Rácz, MHAS
Dr. Géza Györgyi, PhD

Eötvös Loránd University, Faculty of Science

Graduate School in Physics

Head: Dr. Ferenc Csikor, D.Sc

Statistical Physics, Biological Physics and
Physics of Quantum Systems Program

Head: Dr. Jenő Kürti, D.Sc.

Department of Theoretical Physics

Eötvös Loránd University

Budapest, 2012

Introduction

Extreme value statistics (EVS) is an intensively studied field among the applied statistical disciplines. The interest has been growing recently, since the rarely occurring but extreme events, such as earthquakes, hurricanes, floods or droughts, may have great impact on our lives.

The theory of extreme values is well known for independent and identically distributed (iid) variables. If the sample size is going to infinity, the limiting extreme distributions can be classified into three universal classes. The empirically observed extreme distributions, however, can deviate from the limit distributions for several reasons. The convergence towards the limit distributions is typically slow, therefore the corrections due to finite sample size are important. In some cases the maxima are selected from batches of random sizes, and new limit distributions, which depend on the sample size distribution, can be obtained. Furthermore, the variables are often correlated, and the iid theory cannot be applied. In the dissertation we study the above problems, namely the effect of finite and random sample size on extreme statistics, and the effect of correlations on order statistics.

Applications of extreme value statistics require large datasets due to the slow convergence towards the limit distributions. Galaxy samples of recent wide angle surveys are just large enough to attempt an analysis of EVS, and here we study the statistics of maximal luminosity of galaxies observed in a solid angle of the sky. Furthermore, the galaxy luminosity distribution per volume, or the luminosity function, is one of the most basic statistics measured in galaxy surveys, and the large argument asymptotics of this distribution can be ascertained using EVS. The count of galaxies in a solid angle is a finite and random value, therefore the corrections due to finite and random sample size are relevant in the empirical distribution of maximal luminosities. Our theoretical results concerning these corrections have been applied here for the first time.

We examined the magnitude gap between the first and second ranked galaxy as well, since this value is much discussed in astrophysics. Namely, the average gap in a given galaxy cluster is found to be substantially different from what can be obtained from the iid theory. In order to explain this ob-

servation various theories have been worked out assuming a special evolution of the brightest galaxies, and it remains a question whether the difference is caused by dynamical or statistical reasons.

Methods and data analysis

In the dissertation both analytical calculations and large scale simulations have been performed. Furthermore, we carried out statistical analysis of large datasets.

Analytic methods are used to study the EVS in the case of finite and random sample size and to describe the effect of correlations. The finite size correction function is determined by introducing a renormalization group transformation. The effect of correlation in order statistics of $1/f^{\alpha}$ signals is examined with simulations and phenomenological arguments. The average gap between the first and second largest value is obtained analytically for iid variables, as well as the finite size correction of the gap in the Fisher-Tippett-Gumbel class.

We use photometric and spectroscopic data of galaxies from Sloan Digital Sky Survey Data Release 8 (SDSS-DR8) to investigate the statistics of maximal luminosities. The galaxies studied are the Main Galaxy Sample (MGS), divided into MGSblue and MGSred subsamples, as well as the Luminous Red Galaxies (LRGs), at three resolution maps (defined by $N_{side} = 16, 32, 64$) of the HEALPix tessellation on the sky. Selecting carefully the data we eliminate the uncertain and false luminosity values, and we use corrections to get comparable luminosities for the galaxies observed in different redshift intervals. Using numerical simulations we generate larger dataset according to the empirical distributions.

Theses

1. For the Fisher-Tippett-Gumbel (FTG) universality class we determine the exact form of the leading order finite size correction function in the common standardization used in applications (with zero mean and unit standard deviation). Calculating explicitly several parent distributions we show the different functional forms of finite size scaling. We discuss in detail the sample size (N) dependence of the correction for the case of pure exponential (the correction is $\sim 1/N$), Gaussian ($1/\ln N$), e^{-x}/x^{α} type ($1/\ln^2 N$) and lognormal ($1/\sqrt{\ln N}$) distribution respectively.

2. The average gap between the first and second largest values is studied for iid variables. We show that the $N \to \infty$ limiting value depends on the EVS universality class, and for the FTG class we determine the finite size correction, which is an important quantity in astrophysics. Concerning the convergence towards the limiting value we find the scaling to be similar to that observed for the limit distributions of EVS, namely the correction is proportional to $1/\ln N$ for Gaussian type, and $\sim 1/N$ for pure exponential parent distribution.

3. Extreme value statistics is studied for cases, where the sample size (N) is a random variable. Here the limit of infinite sample size can be defined through a typical value (N_0) of the sample size going to infinity. We show analytically that for a given N-distribution three universality classes emerge, according to the asymptotics of the parent. Inside a given class we can further distinguish two families according to the scaling properties of the N-distribution. If the N-distribution scales with N_0, then the limiting distribution of the extremes depends on the scaling function of the N-distribution. The randomness of N does not influence the limiting distribution, if the N-distribution shrinks under the scaling sufficiently fast. For the case of scaling N-distribution we determine the amplitude and shape of the first order finite size correction function using the renormalization group method.

4. Examining the galaxy counts seen in a given solid angle of the sky we show, that the galaxy count distributions of different populations can

be scaled together by the average, and the scaling function can be well fitted with a gamma distribution. We analytically determine the limit distribution of extremes, as well as the finite size correction function for a gamma sample size distribution. Since the scaling function of the sample size distribution is narrow and it shrinks with $\langle N \rangle$, the limit distribution of extreme luminosities turns out to be the FTG distribution, and the effect of random N can be handled as a correction.

5. Analyzing the empirical maximal luminosity distribution of galaxies seen in a given solid angle we find that the EVS of luminosities can be understood within the iid theory, and the empirical distributions are in good agreement with the FTG. In order to draw this conclusion, however, we need to take into account the corrections due to the finite and random sample size.

6. For the magnitude gap between the first- and second brightest galaxies and the Tremaine-Richstone ratios we show the slow, logarithmic N-dependent convergence to the limiting values both analytically and by simulations. This slow convergence hampers the drawing of definite conclusions about the real values of the Tremaine-Richstone ratios for the presently available galaxy samples.

Conclusions

Studying the empirical maximal luminosity distribution of galaxies, our results show that the corrections arising due to finite and random sample size are important, and cannot be neglected in the applications. We can learn also, that even the largest galaxy database at present is not enough for a reliable extreme value statistics. We show that the correct handling of the parent distribution is also important, since its asymptotic behavior strongly affects the theoretical limit distributions and corrections.

Publications underlying the PhD theses

Finite-size scaling in extreme statistics
G. Györgyi, N. R. Moloney, K. Ozogány, and Z. Rácz
Phys. Rev. Lett. 100, 210601 (2008),

Renormalization-group theory for finite-size scaling in extreme statistics
G. Györgyi, N. R. Moloney, K. Ozogány, Z. Rácz, and M. Droz
Phys. Rev. E 81, 041135 (2010),

Order statistics of $1/f^\alpha$ signals
N. R. Moloney, K. Ozogány, and Z. Rácz
Phys. Rev. E 84, 061101 (2011),

Distribution of maximal luminosity of galaxies in the Sloan Digital Sky Survey
M. Taghizadeh-Popp, K. Ozogány, Z. Rácz, E. Regős, and A. Szalay
submitted to ApJ, arXiv:1204.0151 (2012).

Poster related to the PhD theses

Extreme statistics with random sample size and its application on distribution of maximal luminosities
M. Taghizadeh-Popp, K. Ozogány, Z. Rácz, E. Regős, and A. Szalay
poster at the workshop on "Extreme Events: Theory, Observations, Modeling, and Prediction" , Palma de Mallorca, Spain (2008).

Additional publication in the subject of the PhD

Maximal height statistics for $1/f^\alpha$ signals
G. Györgyi, N. R. Moloney, K. Ozogány, and Z. Rácz
Phys. Rev. E 75, 021123 (2007).