Discrete Parameter Estimation for Rare Events:
From Binomial to Extreme Value Distributions

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

im Promotionsprogramm
"PhD School of Mathematical Sciences (SMS)"
der Georg-August University School of Science (GAUSS)

vorgelegt von
Laura Fee Schneider
aus Kassel

Göttingen, 2019

Betreuungsausschuss:

Dr. Andrea Krajina
Rabobank, Utrecht

Prof. Dr. Axel Munk
Institut für Mathematische Stochastik, Universität Göttingen

Prof. Dr. Tatyana Krivobokova
Institut für Mathematische Stochastik, Universität Göttingen

Mitglieder der Prüfungskommission:

Referentin:
Dr. Andrea Krajina
Rabobank, Utrecht

Korreferentin:
Prof. Dr. Tatyana Krivobokova
Institut für Mathematische Stochastik, Universität Göttingen

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Axel Munk
Institut für Mathematische Stochastik, Universität Göttingen

Jun.-Prof. Dr. Madeleine Jotz Lean
Mathematisches Institut, Universität Göttingen

Prof. Dr. Stephan Waack
Institut für Informatik, Universität Göttingen

Dr. Michael Habeck
Institut für Mathematische Stochastik, Universität Göttingen

Tag der mündlichen Prüfung: 26.04.2019

Acknowledgements

At this point, I want to thank my advisory committee for facilitating an interesting and diversified selection of research opportunities. Without their support this dissertation would not have been possible. I am grateful to Dr. Andrea Krajina for being my first supervisor and introducing me to the field of extreme value analysis and its great research community. I express my thanks to Prof. Dr. Axel Munk for the possibility to conduct research in Bayesian statistics and for always providing helpful feedback. I am also thankful to Prof. Dr. Tatyana Krivobokova for her assistance and promotion especially during the last year of my PhD.
Further, I want to thank Prof. Dr. Johannes Schmidt-Hieber for establishing the posterior contraction result with me. I have learned a lot about how to tackle complex theoretical problems. I am thankful as well to my committee, Prof. Dr. Stephan Waack, Jun.-Prof. Dr. Madeleine Jotz Lean and Dr. Michael Habeck, for being part of my disputation. I am also grateful to the DFG for funding my PhD research position in the RTG 2088.
This dissertation benefited from the thoughtful feedback by Thomas Staudt, Dr. Claudia König and Dr. Marco Singer, who were so nice to proof-read it.

I really enjoyed the last years at the IMS, because of the very open and welcoming atmosphere, the many nice colleagues and the new friends I found there. I especially want to thank the *Caphy group* for lovely lunch breaks and Robin Richter, Atika Batool, Dr. Gabriel Berzunza Ojeda, Dr. Marco Singer and Peter Kramelinger for productive offices atmospheres throughout the years.
Finally and most importantly, I owe many thanks to my family and friends, especially to Patrick Nern for always being there and supporting me, to my parents Uwe and Constanze Schneider for always encouraging and empowering me and to Nadine Velte, Finn Schneider, Stefanie Best, Runa Förster and Katharina Breininger for providing little distractions from work.

Preface

Estimating a discrete parameter from rare events is a challenging task, since observations are scarce in such situations. In this cumulative dissertation two distinct problems resulting from rare events are discussed and methodologies to solve them are suggested. First, we employ a Bayesian approach in the binomial model to overcome a lack of information on the parameter n that arises from a small success probability. In this demanding setting we derive a posterior consistency statement that delivers a clearer theoretical understanding for the asymptotic behaviour of Bayesian estimators. Secondly, we statistically investigate events in the tail of heavy-tailed distributions. For this task, the peak-over-threshold approach is a common model, which crucially depends on the selection of a high threshold above which observations can be used for statistical inference. To improve the utility of threshold selection procedures, we propose two new methods and evaluate their performance theoretically and numerically in comparison to other approaches.

The dissertation is based on three publications, which are listed in the addenda and can be found in chapters A, B and C. The articles Schneider et al. (2018a) and Schneider et al. (2018b), which address the binomial problem, can be found in Chapters A and B. Chapter C comprises the article Schneider et al. (2019) on threshold selection in extreme value analysis.

Chapter 1 provides an overview of the challenges related to rare events and introduces the specific scenarios that we study in more detail in the following chapters.

Chapter 2 overviews the work on posterior consistency in the binomial model in publications A and B. In Section 2.1 a comprehensive discussion is presented about the existing literature on estimating the binomial parameter n when p is unknown. The main contributions of the two manuscripts are explained in more detail in Section 2.2. A discussion and an outlook on these results are presented in Section 2.3, and my own contribution to the articles is pointed out in Section 2.4.

In Chapter 3 the contributions of publication C are outlined and assessed. Necessary background knowledge about threshold selection in extreme value analysis is summa-

rized in Section 3.1. In Section 3.2 we elaborate on our novel methods and results, while interesting aspects for future research are discussed in Section 3.3. My own contribution to article C is clarified in Section 3.4.

Contents

List of Symbols

$\mathbb{R}$	Set of real numbers
$\mathbb{N}$	Set of positive integers
$\lvert\cdot\rvert$	Absolute value
$n_k = O(a_k)$	$\exists\, K, M > 0 :\ \lvert n_k\rvert \leq M a_k,\ \forall k \geq K$
$n_k = o(a_k)$	$\lim_{k\to\infty} \lvert n_k / a_k\rvert = 0$
$\xrightarrow{\mathscr{D}}$	Convergence in distribution
$\xrightarrow{\mathbb{P}}$	Convergence in probability
$X \sim F$	Random variable X is distributed following the distribution function F
$X \mid Y$	X conditioned on Y
$\mathbb{E}[X]$	Expectation of X
$\mathrm{Var}(X)$	Variance of X
$\mathrm{Cov}(X, Y)$	Covariance of X and Y
$X_k = O_p(a_k)$	$\forall \epsilon > 0\, \exists\, M > 0, K > 0 :\ \mathbb{P}(\lvert X_k / a_k\rvert > M) < \epsilon,\ \forall k > K$
$X_k = o_p(a_k)$	$\lim_{k\to\infty} \mathbb{P}(\lvert X_k / a_k\rvert \geq \epsilon) = 0,\ \forall \epsilon > 0.$
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\mathrm{Bin}(n, p)$	Binomial distribution with $n \in \mathbb{N}$ and success probability $p \in [0, 1]$

CHAPTER 1

Introduction

Statistical inference on rare events is highly challenging, since only little information is available about the features of interest. Often especially these properties, which are difficult to perceive, are of major concern in applications. For example, it is very important for an insurance company to know about the risk of extreme events such as violent wind storms, which can cause the highest losses, but are rarely or not at all observed. Another example is estimating the population size of a species from random counts, where the species is very cautious and therefore scarcely sighted.

The first example considers extreme events lying in the tail of heavy-tailed distributions and is investigated by the field of extreme value analysis. The second example can be modelled by a binomial distribution with a small and unknown success probability p and the population size of the species corresponds to the discrete parameter n to be estimated. In both cases one is interested in the tail of the distribution, but the observations mainly lie in the bulk of the distribution around its expectation. In Figure 1.1, an exemplary histogram of a binomial sample with a small success probability is presented. The parameter $n = 20$ is the right endpoint of the distribution and is supposed to be estimated from the data, but the maximal observation is only 3. This strongly illustrates the difficulties related to the estimation task.

The key point when handling rare events is that one needs to take advantage of a priori

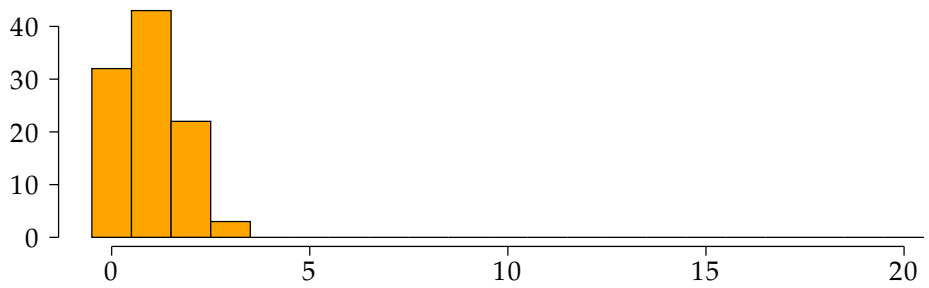


Figure 1.1: Histogram of a binomial sample of size 100 with $n = 20$ and $p = 0.05$.

known structural properties in order to improve the quality of estimation. Completely non-parametric approaches are likely to fail on scarce observations.

In case of the binomial distribution we already have a parametric model and still need to overcome the lack of information about n and p. This is analysed in terms of a Bayesian approach in article A. We theoretically investigate this problem via a posterior contraction statement and then tackle it in an application in super-resolution microscopy by including prior information about p. In contribution B we gain further insights about the asymptotic behaviour beyond the theoretical result via an extensive numerical study. For studying extreme values we utilize the peak-over-threshold approach, which is based on the limiting distribution of observations above a high threshold. The asymptotic theory provides extrapolation from large observations further into the tail to even more extreme events. For statistical inference it is necessary to select a suitable threshold, and for this task we suggest new procedures in article C.

First we have a closer look at the problem of estimating the discrete parameter n of the binomial distribution. The binomial distribution describes the number of successful outcomes of n Bernoulli trials each with success probability p, i.e., if $X \sim \mathrm{Bin}(n, p)$ with $n \in \mathbb{N}$ and $p \in [0, 1]$,

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, \dots, n.$$

In article A we suggest a Bayesian approach for estimating the binomial parameter n from independent identically distributed (i.i.d.) observations, and we prove posterior consistency for rare events, meaning that we let $p \to 0$ and $n \to \infty$ as the number of observations grows. Our motivation for studying the binomial distribution with small success probability p comes from an application in quantitative nanoscopy. There, the total number of fluorescent markers (fluorophores) attached to so-called DNA-origami is estimated from a time series of microscopic images. The number of active fluorophores counted at each DNA-origami is modelled as binomial observation, where the probability p that a fluorophore is active in each image is very small (often below 5%). In Figure 1.2, two such microscopic images are displayed. Each frame is recorded at a different time point t and contains a number of binomial observations $X_i^{(t)}$. These observations are the number of active fluorophores at a specific DNA-origami on the frame, which can be determined from the brightness at this spot.

This setting, where the success probability p is small (and n potentially large), is very challenging. Additionally to the histogram in Figure 1.1 the arising difficulties can be understood by considering the following property of the binomial distribution: if

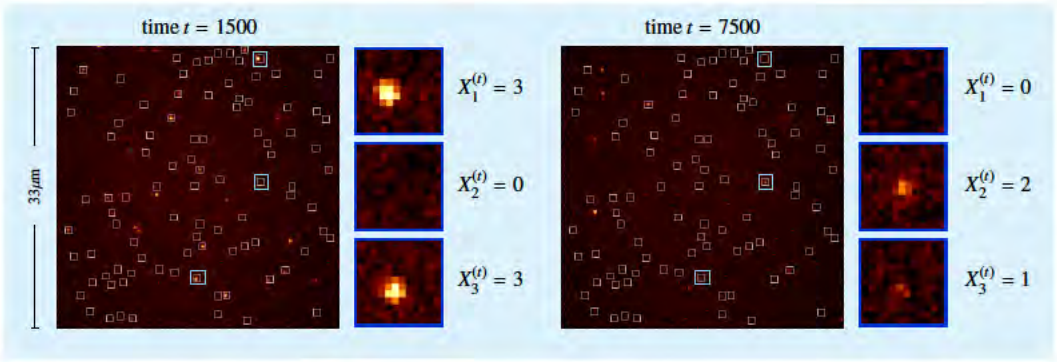


Figure 1.2: Fluorescence microscopy images at two time points with magnified pictures of three exemplary regions of interest. The regions of interest specify the location of DNA-origamis in the images. In each region the brightness represents the binomial observation.

n converges to infinity, p converges to zero, and the product np converges to $\lambda > 0$, then a $\text{Bin}(n, p)$ random variable converges in distribution to a Poisson variable with parameter λ. Thus, the binomial distribution converges to a distribution with a single parameter. This suggests that it gets harder to derive information about the two parameters separately when n becomes large and p small.

Estimating n in this situation is more manageable if one includes prior information about p. We therefore consider a Bayesian approach to estimate n. In order to understand the performance of the estimator in the previously described challenging scenario of rare events without additional knowledge of p, we establish posterior consistency in an asymptotic setting where $n \to \infty$, $p \to 0$ and $np \to \lambda$ as the sample size k increases. We show that $n^{6+\epsilon} = O(k)$ for $\epsilon > 0$ is a sufficient bound on the growth of n to ensure that consistency of the estimator still holds. This result for Bayes estimators is to our knowledge the first statement that goes beyond consistency for fixed parameters and it is especially interesting when compared to the sample maximum. The sample maximum is a consistent estimator for n and converges exponentially fast to the true value if n and p are fixed, as described in the introduction of publication A. However, DasGupta and Rubin (2005) discuss that it is often not a useful estimator in practice. We theoretically substantiate this fact, since the maximum needs a much larger sample size relative to n to still be consistent: in the asymptotic scenario, where $n \to \infty$ and $p \to 0$, the bound $n^n = O(k)$ is necessary for consistency.

We conduct a simulation study in article B to further investigate the bound $n^{6+\epsilon} = O(k)$ numerically. We observe that the theoretical result is probably slightly too strict and a relaxed bound $n^\alpha = O(k)$, where $\alpha \approx 4$, seems possible.

The second field of interest is extreme value analysis, which enables estimating the risk of extreme and rare events. Here we are concerned with heavy-tailed distributions and the peak-over-threshold method, as compared to the block maxima approach (see e.g. Chapter 1 in Dey and Yan (2016)). Therefore, we consider i.i.d. random variables $X_1, \ldots, X_n$ with distribution function F, where n is now the sample size and F is in the domain of attraction of an extreme value distribution G_γ with extreme value index $\gamma > 0$, i.e. there exists $a_n > 0$ and $b_n \in \mathbb{R}$ such that for $x > 0$

$$F^n(a_n x + b_n) \longrightarrow G_\gamma(x) = \exp\left(-x^{-1/\gamma}\right), \quad \text{as } n \to \infty.$$

In this situation the conditional random variable $(\log(X_1/t) \,|\, X_1 > t)$ converges in distribution to an exponential random variable with expectation γ, as $t \to \infty$. The peak-over-threshold approach utilizes this limiting distribution to approximate the logarithm of the observations above a high threshold t with the exponential distribution. The selection of the threshold t above which the data can be approximated by an exponential distribution and used for statistical inference about the tail is one of the most fundamental problems in this field of statistics. It is common to let t be the $(n-k)$-th order statistic and choose the discrete sample fraction k of largest observations instead of the threshold t. We focus on this problem of selecting k and highlight its crucial influence on statistical results for the exemplary task of estimating the extreme value index γ. To estimate γ we employ the Hill estimator (Hill, 1975), which is the mean of the rescaled exceedances of the threshold,

$$\hat{\gamma}_k := \frac{1}{k} \sum_{i=1}^{k} \log\left(X_{(n-k+i,n)}\right) - \log\left(X_{(n-k,n)}\right), \tag{1.1}$$

where $X_{(1,n)} \leq \cdots \leq X_{(n,n)}$ denote the order statistics of a sample of size n. Figure 1.3 presents the Hill estimator as a function in k for three Fréchet samples of size 500 and illustrates the critical influence of the sample fraction k. These Hill plots demonstrate

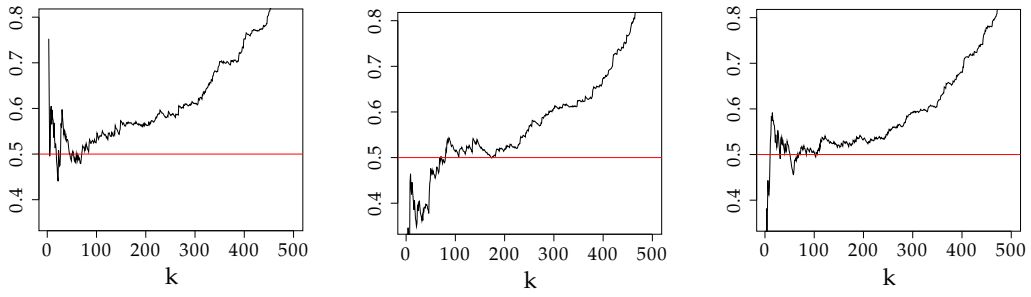


Figure 1.3: Hill plots of three Fréchet(2) samples with extreme value index $\gamma = 0.5$.

some main characteristics of the estimator and the peak-over-threshold approach in general. If k is small, only few observations are used and the estimator is highly varying. On the other hand, if k becomes larger and the threshold therefore smaller, the exponential approximation does not hold anymore and a bias occurs. The true value of γ is marked by the dotted line and only achieved by a few values of k. Thus, selecting k is hard and can be considered a classical problem of bias-variance trade-off. The sample fraction k that solves this problem and minimizes the asymptotic mean square error (AMSE) of the Hill estimator is often considered the optimal sample fraction k_{opt} and targeted by many threshold selection procedures.

In publication C we introduce two new procedures for selecting the sample fraction. The first approach is based on measuring the deviation of the data from the exponential approximation. We construct an error functional denoted as inverse Hill statistic (IHS), which aims at estimating an integrated square error of the exponential density under the hypothesis that the observations $\log(X_{(n-k+i,n)}) - \log(X_{(n-k,n)})$ in (1.1) are indeed exponentially distributed. This method is easy to compute and does not depend on any tuning parameters. We can improve its performance via smoothing IHS and subsequently minimize it to select k. The smoothed IHS performs remarkably well in the simulation study when utilized for adaptive quantile estimation.
In the second approach we smoothly estimate the AMSE of the Hill estimator. By minimizing the estimator called SAMSEE (smooth AMSE estimator) we estimate k_{opt}. The idea for constructing SAMSEE is based on obtaining a preliminary estimate for γ and the bias of the Hill estimator. For the extreme value index we apply the generalized Jackknife estimator (Gomes et al., 2001), and for the bias component we use the difference between averages of the Hill estimator. This way, we obtain an estimator for the AMSE, which only depends on one tuning parameter K. We further suggest a selection procedure for K, which then enables a completely automated selection of the sample fraction. In our simulation study SAMSEE performs very stable and is on average over all the considered distributions superior to other methods estimating k_{opt}. Both methods are further employed in an application for the local selection of the threshold in order to estimate an extreme value index varying over time. This example highlights some advantages of the suggested procedures and the benefit of completely data-driven threshold selection approaches.

CHAPTER 2

Posterior Consistency in the Binomial Model

This chapter evaluates the two articles A and B. It offers a broad literature review on estimating the binomial parameter n in Section 2.1 and a discussion of the main results of the publications A and B in Section 2.2. The impact of these articles on possible future research is presented in an outlook in Section 2.3. Finally, in Section 2.4 my own contributions to these articles are pointed out.

2.1 Literature Review

Estimating the discrete parameter n of the binomial distribution has been considered multiple times over the last decades, starting with Haldane (1941) and Fisher (1941). Fisher refers to the problem as purely academic and not very interesting, since n is the maximum of a sufficiently large sample. This is correct in theory, as the maximum converges exponentially fast to the true value n in probability. However, this fact is not very useful in practice, as analysed and explained by DasGupta and Rubin (2005). They illustrate that the sample size often has to be infeasibly large to obtain an estimate close to n with high probability. For the parameter setting $n = 20$ and $p = 0.1$ for example, more than 900,000 observations are necessary to guarantee that the maximum is larger than $n/2$ with probability greater than $1/2$. In their paper, DasGupta and Rubin (2005) also present two new estimators for n, where one is a bias corrected version of the sample maximum and the other is a novel moment estimator.

Olkin et al. (1981) analyse some problems with the classical maximum likelihood estimator (MLE) and the method of moments estimator (MME) for n, which occur if the success probability is close to zero. More precisely, they prove that the MLE can become infinite and the MME negative if p is small, and they suggest stabilized versions of these classical estimators. This way, issues arising from rare events are considered but not studied asymptotically in terms of consistency or depending on parameter sequences.

Another estimator for n is the Carroll-Lombard (CL) estimator (Carroll and Lombard, 1985), which is also a stabilized MLE. The idea is to assume a beta prior on p and maximize the marginal likelihood of n, the beta-binomial likelihood. This estimator turns out to be much more stable than the MLE. The CL-estimator can be interpreted as a Bayesian maximum a-posteriori (MAP) estimator with constant prior on n. Several Bayesian approaches have been suggested for the binomial distribution starting with Draper and Guttman (1971) and followed by e.g. Raftery (1988), Günel and Chilko (1989) and Hamedani and Walter (1988). All of these models are based on p being beta distributed. This seems to be a natural choice, as it is the conjugate prior for the binomial distribution. The mentioned suggestions differ in the choice of the prior on n and the specific loss function. Draper and Guttman (1971) use a uniform prior for n on $\{1, \dots, N_0\}$ with $N_0 \in \mathbb{N}$ and the MAP estimator, which minimizes the 0/1-loss. Note that this proposal only deviates from the Carroll-Lombard estimator by the upper bound N_0 and thus gives the same estimates if N_0 is large enough. A hierarchical Bayes approach is introduced by Raftery (1988), utilizing a Poisson prior on n and the relative quadratic loss function.
While it is standard to use a beta prior on p, there is extensive discussion on the best prior choice for n. Kahn (1987), for example, cautions against the use of the improper constant prior distribution, which can lead to improper posteriors. Link (2013) also considers this problem and compares the uniform prior to the scale prior, which is proportional to $1/n$. Berger et al. (2012) and Villa and Walker (2014) discuss possible constructions of objective prior distributions for n, since standard procedures using the Fisher information can not be employed for a discrete parameter.

For all previously mentioned Bayes estimators, posterior consistency for fixed parameters n and p follows directly from the Theorem by Doob, see e.g. van der Vaart (1998) pp.149. No further theoretical results are known to the author.
For frequentist estimators, it is also not clear how to derive standard asymptotic results beyond consistency. It seems common to let n grow to infinity with the sample size and consider the relative error of the estimator. In Carroll and Lombard (1985) they show asymptotic normality for their CL-estimator $\hat{n}_{CL}$ in the situation that p is constant, $n \to \infty$ and $\sqrt{k}/n \to 0$ as $k \to \infty$. Then it holds that

$$\sqrt{k}\left(\frac{\hat{n}_{CL}}{n} - 1\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, 2(1-p)^2/p^2\right). \tag{2.1}$$

Under the same considerations Blumenthal and Dahiya (1981) study the MME and MLE. These statements address a situation that is very different from ours, where

$n \to \infty$ and $p \to 0$. In case of a fixed p and $n \to \infty$, a properly rescaled binomial distribution converges to a normal distribution. The normal limit simplifies the estimation of n considerably when compared to the difficulties arising from the Poisson limit distribution. The analysis of the estimators in DasGupta and Rubin (2005) is an exception, because n stays constant and they do not consider the relative difference.
The theoretical study by Hall (1994) provides an analysis of the CL-estimator, the MLE and the MME in an asymptotic scenario which is closely related to our set-up. Hall also considers that $n \to \infty$, $p \to 0$ and $np \to \mu \in (0, \infty)$ with $k \to \infty$. However, he additionally needs a lower bound on the rates and studies the relative error, which is rescaled with n, whereas we consider $\hat{n} = n$ instead of $\hat{n}/n = 1$. Due to these differences, the approaches are not directly comparable.

2.2 Main Results

As discussed in the introduction, estimating the binomial parameter n from scarce observations is challenging. This motivates us to consider a Bayesian estimator, since it provides a natural way of including additional knowledge about p. We consider the binomial parameters to be random variables, N and P, following specific distributions. We let P be Beta(a, b) distributed with $a, b > 0$, which is a natural prior choice adopted by many authors before, see Section 2.1. Further we assume that N and P are independent, which we also share with the afore mentioned approaches. In this scenario we provide three contributions to the problem of estimating n from rare events. We start with introducing a new class of Bayes estimators and then prove posterior consistency in an asymptotic scenario, where $n \to \infty$ and $p \to 0$, to theoretically understand arising difficulties more thoroughly. Lastly, we complement the theory by a numerical study to investigate the tightness of the asymptotic bounds.

For the construction of the estimator, we define the following class of prior distributions Π_N on N,

$$\Pi_N(n) \propto n^{-\nu}, \text{ for } \nu > 1,$$

denoted as scale priors. If the scale parameter is $0 \leq \nu \leq 1$, Π_N is an improper prior, since it is no longer a probability distribution. An improper prior can still lead to a proper posterior distribution and can be useful in applications. For this reason, we also include improper priors in our theoretical analysis and application in publication A.
For the loss function we follow the arguments in Raftery (1988) and use the relative

quadratic loss, $l(x, y) = (x/y - 1)^2$. This leads to the definition of our scale estimators,

$$\hat{n} := \frac{\mathbb{E}[N^{-1}|\mathbf{X}^k]}{\mathbb{E}[N^{-2}|\mathbf{X}^k]} = \frac{\sum_{n=X_{(k,k)}}^{\infty} n^{-(1+\nu)} L_{a,b}(n)}{\sum_{n=X_{(k,k)}}^{\infty} n^{-(2+\nu)} L_{a,b}(n)}, \tag{2.2}$$

where $\mathbf{X}^k$ denotes the binomial i.i.d. sample of size k and $L_{a,b}$ denotes the beta-binomial likelihood, see e.g. Carroll and Lombard (1985). It is possible to insert information about p through the choice of the parameters a and b of the beta prior. In the simulation study in article A we come to the conclusion that in most cases it is beneficial to choose values like $a = 2$, which lead to a unimodal but not very concentrated density. The parameter b should then be selected in such a way that the expectation of the prior equals a preliminary expected value $\tilde{p}$ of p, i.e. $b = a(1 - \tilde{p})/\tilde{p}$. In the application in article A, for example, such a provisional estimate $\tilde{p}$ is obtained from a second experiment. This enables us to increase the efficiency in estimating n in the case of fluorophore counts.

In the theoretical analysis, we study posterior contraction in an asymptotic scenario describing rare events. We assume the independence between P and N and that P is beta distributed, but only a tail bound is necessary to hold for the prior on N,

$$\Pi_N(n) \geq c_2 e^{-c_1 n^2}, \ n \in \mathbb{N},$$

for some positive constants c_1 and c_2. This covers the scale estimator (2.2) but also many other estimators. In this model we can show that posterior consistency holds for the following set of parameter sequences, where the parameters may depend on the sample size k,

$$\mathcal{M}_\lambda := \left\{(n_k, p_k)_k : 1/\lambda \leq n_k p_k \leq \lambda, \ n_k \leq \lambda \sqrt[6]{k/\log(k)}\right\},$$

for $\lambda > 1$. The set $\mathcal{M}_\lambda$ greatly generalizes the scenario of fixed parameters, as it also allows for sequences with $n \to \infty$ and $p \to 0$ at certain rates. In Theorem 1 in publication A we prove that

$$\sup_{(n_k^0, p_k^0) \in \mathcal{M}_\lambda} \mathbb{E}_{n_k^0, p_k^0}\Big[\Pi(N \neq n_k^0 \,|\, \mathbf{X}^k)\Big] \to 0, \text{ as } k \to \infty, \tag{2.3}$$

where $X_1, \ldots, X_k \overset{i.i.d.}{\sim} \text{Bin}(n_k^0, p_k^0)$ and $\Pi(\ \cdot\ |\, \mathbf{X}^k)$ is the marginal posterior probability of N given the sample $\mathbf{X}^k$. This theorem extends posterior consistency for fixed parameters to the situation of rare events. From result (2.3) follows directly the consistency of the scale estimators in (2.2) as well as the consistency of many of the estimators mentioned

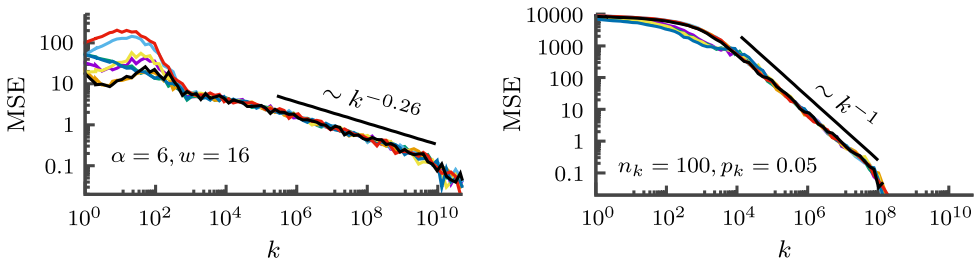


Figure 2.1: Log-log plots of the MSE of various Bayesian estimators for the binomial parameter n_k as a function in the sample size k. In the left plot the parameters are $n_k = wk^{1/\alpha}$ and $p_k = \mu/n_k$ with $\alpha = 6$, $w = 16$ and $\mu = 25$, and on the right the parameters are fixed to $n_k = 100$ and $p_k = 0.05$. The dashed lines indicate the sub-asymptotic convergences rates.

in Section 2.1 for parameter sequences in $\mathcal{M}_\lambda$. This is discussed in detail in article A.

Article B numerically investigates the asymptotic behaviour of the Bayes estimators to evaluate the tightness of the bound on n_k in $\mathcal{M}_\lambda$, which ensures consistency. The simulations strengthen our theoretical result, as they indicate the optimality of a polynomial rate, but they also emphasize a slightly too conservative value of the exponent. The numerical study suggests that the theoretical bound $n^{6+\epsilon} = O(k)$ with $\epsilon > 0$ can possibly be relaxed to $n^\alpha = O(k)$, where $\alpha \approx 4$.

The article also provides further insight about the speed of convergence of the estimators depending on α if $n \propto k^{1/\alpha}$. We define sequences $n_k = wk^{1/\alpha}$ and $p_k = \mu/n_k$ with $\alpha, \mu, w > 0$ and simulate the mean square error (MSE) of various Bayesian estimators for increasing samples of size k. This way, we complement our previous results by describing the rates of convergence of the MSE that become visible as a linear segment in Figure 2.1 and that are denoted as sub-asymptotic convergence rates. The log-log plots illustrate that the statement of Theorem 1 holds and the estimators are consistent for $\alpha = 6$, even though the rate at which the MSE decreases appears to be much slower than for constant parameters. One can observe that the MSE of the estimators decreases at a faster and probably exponential rate if the error is already very small, but the rate before this point seems more interesting, since it appears in a regime which is more relevant for applications. We have seen before that the sample maximum converges exponentially fast for fixed n and p, but it does not perform well in practice. This is highlighted by the simulation study suggesting that the sample maximum converges much slower to the true value n. It seems that the MSE of the sample maximum converges with a rate of $k^{-0.13}$ in the sub-asymptotic regime in contrast to the standard parametric rate k^{-1}, which we observe for the Bayesian approaches in Figure 2.1.

2.3 Discussion and Outlook

The two publications A and B deliver interesting starting points and incentives for further research about Bayesian inference on the binomial distribution.
To our knowledge, statement (2.3) is the only asymptotic result for a Bayesian approach on the binomial problem that goes beyond posterior consistency for fixed parameters, which already follows from Doob's theorem, see e.g. van der Vaart (1998). There are not many tools to derive such theoretical results for discrete Bayes estimators and thus no standard approaches. We had to develop a new strategy of proof for our result, and simulations suggest that the obtained bounds are good but not optimal. It seems necessary to further evolve and refine the tools for these asymptotic problems.
Since our motivation for the Bayesian approach is including prior knowledge about p, a natural extension of the main result would be to employ an empirical prior distribution on p. It could be instructive to see how much improvement in the theoretical rates is possible by considering a beta prior which contracts around p. In the application we observe the positive effect of an empirical prior, but such a result would reveal more concrete information about the benefit of prior knowledge.
Additionally, it would be interesting to theoretically analyse the sub-asymptotic behaviour of the estimators observed in the numerical study in publication B. The exponentially fast convergence only kicks in when the MSE is already very small. Thus, the convergence rate of interest is the sub-asymptotic rate before the exponential decay. Deriving these sub-asymptotic convergence rates theoretically could be helpful to make informed decisions about optimal estimators or necessary sample sizes.

2.4 Own Contribution

Contribution to A: My main contribution to this paper is the posterior contraction result of Theorem 1. I developed the theoretical framework jointly with Johannes Schmidt-Hieber and then elaborated the mathematical details. I finalized additional statements for the proof with the assistance of Thomas Staudt. Further, I wrote in most parts the theory and simulations sections. Overall the paper was formulated jointly with the co-authors.

Contribution to B: My main contribution for this article was coming up with the idea for the simulation study and writing the manuscript, where I benefited from helpful comments by Thomas Staudt and Axel Munk. The simulation study was implemented and visualized by Thomas Staudt.

CHAPTER 3

Threshold Selection in Extreme Value Analysis

This chapter offers a close look at the contribution C. Section 3.1 provides an overview of existing literature about threshold selection in univariate extreme value analysis, whereas in Section 3.2 the novel procedures are summarized. An outlook and discussion of the results and future research are offered in Section 3.3. In Section 3.4 my own contribution to the article is explained.

3.1 Extreme Value Analysis Review

Extreme value theory investigates very rare events and consists of probabilistic results about the tail of a distribution. The foundation was laid by the early works of Fisher and Tippett (1928) and Gnedenko (1943). They proved that the non-degenerate limit of a rescaled maximum of i.i.d. observations lies in the class of extreme value distributions $G_\gamma(x) := \exp\{-(1+\gamma x)^{-1/\gamma}\}$ for $\gamma \in \mathbb{R}$ and $1+\gamma x > 0$. If a distribution F possesses such a limit, it belongs to the domain of attraction of G_γ. The shape parameter γ is called the extreme value index. It describes if a distribution has a finite right endpoint ($\gamma < 0$), an exponential decay in the tail ($\gamma = 0$) or is heavy-tailed ($\gamma > 0$). Extreme value distributions provide a possibility to extrapolate from the data further into the tail via the use of block maxima. The idea is to divide observations into blocks and then take the maximal observation within each block. These maxima are then used to estimate γ as well as the shift and scale parameter of the extreme value distribution.
A further important theoretical result is the Pickands-Balkema-de Haan theorem derived by Balkema and de Haan (1974) and Pickands (1975). They consider the conditional distribution above a high threshold t and derive a generalized Pareto limiting distribution. The shape parameter of the generalized Pareto distribution is again the extreme value index γ. This leads to the peak-over-threshold methodology, which approximates observations above a high threshold by a generalized Pareto distribution, see Davison and Smith (1990). We concentrate on the peak-over-threshold approach for univariate

heavy tailed distributions. Thus, we consider i.i.d. random variables $X_1, \dots, X_n \sim F$, where F is in the domain of attraction of an extreme value distribution with $\gamma > 0$. Then it holds that

$$\frac{X_1}{t} \,\Big|\, X_1 > t \xrightarrow{\mathcal{D}} P, \text{ as } t \to \infty \text{ and } P \sim \text{Pareto}\left(1, \frac{1}{\gamma}\right),$$
$$\log\left(\frac{X_1}{t}\right) \,\Big|\, X_1 > t \xrightarrow{\mathcal{D}} E, \text{ as } t \to \infty \text{ and } E \sim \text{Exp}\left(\frac{1}{\gamma}\right).$$

These limit relations justify to model observations above a high threshold t with a Pareto or exponential distribution. Compared to the block maxima approach, the peak-over-threshold method facilitates the use of more data. For both methods, however, a bias-variance trade-off occurs. In the former, the blocks need to be large enough, such that the maxima are approximately independent and follow an extreme value distribution, but small enough to provide enough observed maxima. In the latter, a lower threshold reduces the variance by providing more observations, but leads to a higher bias for violating the limit approximation. There are various possible approaches to measure the goodness of the approximation above the threshold, and their optimality depends strongly on the feature of interest and the chosen method or estimator. All these difficulties make threshold selection an intricate problem in extreme value analysis.

In publication C, we tackle the threshold selection problem and search for an appropriate number k of largest observations to be used for statistical inference about the tail. This is equivalent to letting the threshold be the $(n-k)$-th order statistic, meaning $t = X_{(n-k,n)}$. Most existing approaches for this task heavily depend on tuning parameters, which can have a critical influence on the estimate and are hard to interpret. The optimal choice of these parameters usually depends on the underlying unknown distribution, and there is only numerical justification for specific suggestions. Our motivation for studying this problem is mainly to reduce the burden of tuning parameter selection and to make automated data-driven procedures easily available. At this point, we want to present a brief guide through different existing methodologies for selecting the threshold and their associated tuning parameters. More extensive reviews on such procedures are provided e.g. in Scarrott and MacDonald (2012), in Chapter 4 in Dey and Yan (2016) and in Section 4.7 in Beirlant et al. (2004).

We start with mentioning rules of thumb and heuristic approaches. It has been suggested to use $k = \sqrt{n}$ (Ferreira et al., 2003) or the upper 10% of the data (DuMouchel, 1983), which is a popular choice in applications. The use of data visualisation tools is discussed in Kratz and Resnick (1996) and Drees et al. (2000), where one example is the mean residual life plot introduced in Davison and Smith (1990). These graphical diagnostics

are subjective regarding the interpretation of the plot and require the manual choice of a threshold. There are also heuristically motivated techniques trying to automate the interpretation of such graphical diagnostics, for example the method in Reiss and Thomas (2007) that searches for a region of stability among γ estimates. This estimator depends on a tuning parameter whose choice is numerically further analysed in Neves and Fraga Alves (2004).
Another possibility for selecting the threshold are testing approaches. Hill (1975) suggests to employ tests for exponentiality in order to choose a sample fraction. However, Hall and Welsh (1985) show that this tends to overestimate the optimal tail fraction. Goegebeur et al. (2008) extend the work by Hill and suggest a family of kernel based test statistics. A different approach is to consider goodness-of-fit tests comparing the empirical distribution function to the estimated generalized Pareto distribution, see e.g. Bader et al. (2018). It is important to note that all of these tests depend on the choice of the significance level and the corresponding quantile of the test statistic. A lower significance level directly translates into selecting a larger sample fraction.
A related idea is to minimize the distance between empirical and fitted generalized Pareto distribution, as in Pickands (1975), Gonzalo and Olmo (2004) and Clauset et al. (2009). The method in Clauset et al. (2009), for example, selects the sample fraction that minimizes the Kolmogorov-Smirnov distance and is theoretically analysed in Drees et al. (2018). They prove that this approach does not select the asymptotically optimal sample fraction for the Hill estimator in (1.1).
This theoretical consideration leads us to methods that are motivated by choosing the optimal sample fraction for specific estimators. Examples are the bootstrap approaches by Ferreira et al. (2003) for quantile estimation and for estimating tail probabilities by Hall and Weissman (1997) or the optimal bias-robust estimation approach in Dupuis (1998). These procedures are concerned with different estimators and crucially depend on the choice of tuning parameters.

A more frequently considered error functional is the asymptotic mean square error (AMSE) of the Hill estimator (Hill, 1975). The Hill estimator defined in equation (1.1) is a well known estimator for γ and commonly used, despite its high sensitivity to the choice of the sample fraction k. As the mean value of the logarithmic exceedances of the $(n-k)$-th order statistic it estimates the expectation γ consistently if the exponential approximation holds. For this reason, minimizing the AMSE can be understood as a goodness-of-fit measure for the exponential model, as it assesses the deviation of the empirical mean to the true limiting expectation. A semi-parametric formulation of the AMSE can be obtained from the normal limit distribution of the Hill estimator. We

consider the first and second order condition

$$\lim_{t\to\infty}\frac{1-F(tx)}{1-F(t)} = x^{-1/\gamma} \quad \text{and} \quad \lim_{t\to\infty}\frac{\frac{1-F(tx)}{1-F(t)} - x^{-1/\gamma}}{A\left(\frac{1}{1-F(t)}\right)} = x^{-1/\gamma}\,\frac{x^{\rho/\gamma}-1}{\gamma\rho}$$

with the second order parameter $\rho < 0$ and a positive or negative function A, s.t. $\lim_{t\to\infty} A(t) = 0$. Then, if $k \to \infty$, $k/n \to 0$ and $\sqrt{k}A(n/k) \to \lambda$ as $n \to \infty$, asymptotic normality holds, i.e.,

$$\sqrt{k}\,(\hat{\gamma}_k - \gamma) \xrightarrow{\mathscr{D}} \mathcal{N}\left(\lambda/(1-\rho)^2, \gamma^2\right),$$

see Theorem 3.2.5 in de Haan and Ferreira (2006). This leads to the AMSE,

$$\mathbb{AE}[(\hat{\gamma}_k - \gamma)^2] = \gamma^2/k + A(n/k)^2/(1-\rho)^2, \tag{3.1}$$

where $\mathbb{AE}$ denotes the asymptotic expectation. There exist several publications providing procedures and theoretical analysis about estimating the minimizing fraction of (3.1) k_{opt}. Drees and Kaufmann (1998), for example, utilize the Lepskii method to derive estimates for this optimal sequence. Their approach crucially depends on tuning parameters and initial estimates of γ and ρ. The procedure of Guillou and Hall (2001) employs a statistic of accumulated log-spacings and requires the choice of a critical value to test it against. Danielsson et al. (2001) suggest a double bootstrap method extending the previous ideas for a resampling approach by Hall (1990). They need to choose the size of the bootstrap samples for which they introduce a separate selection tool. This additional tool can become computationally very expensive if estimating from large samples. Beirlant et al. (2002) estimate a parametric representation of k_{opt} employing least squares estimates from an exponential regression approach. Their method requires an estimate for ρ and a further tuning parameter. To simplify the use of their approach they consider the median of estimated sample fractions over a range of values of the tuning parameter. In Goegebeur et al. (2008), the properties of a test statistic are used to estimate the bias and construct an estimator for the AMSE/γ and minimize it. This approach is an exception among the afore mentioned ones, as it does not require the choice of a further tuning parameter if $\rho = -1$ is fixed. On the other hand, all other described procedures for estimating k_{opt} offer consistency statements, but their finite sample performance is strongly influenced by more or less intricate tuning parameters.

3.2 Main results

Since the selection of additional tuning parameters can be a burden, as mentioned in Section 3.1, rules of thumb are frequently used in practical applications. Article C provides two new procedures to select the sample fraction k in the peak-over-threshold model, which depend on no or only one tuning parameter and are therefore easier to employ. For the one necessary parameter we offer a stable selection routine. We emphasize that such data-driven and automated procedures can clearly improve the performance of the adaptive estimation. For example, locally selected thresholds can decrease the variance in estimating a time dependent extreme value index. We consider this task in article C and utilize it for analysis of the tail behaviour of operational losses depending on the time as a covariate.

For the first approach we look at the integrated square error (ISE) of the exponential density to its parametric estimator employing the Hill estimator. We assume that the logarithmic exceedances of a high threshold are indeed exponentially distributed and estimate the expectation of ISE under this hypothesis. In this way, we derive the inverse Hill statistic,

$$\mathrm{IHS}(k) := \frac{4-k}{2\hat{\gamma}_k k}.$$

IHS can be highly varying for small k, and therefore we smooth it by a technique taking into account the dependence between the estimates, see e.g. Serra et al. (2018). Minimizing IHS leads to a more conservative choice of the sample fraction than minimizing the AMSE of the Hill estimator, as IHS controls the exponential approximation more strictly. The selected sample fraction is asymptotically smaller than k_{opt}, but brings benefits when estimating high quantiles from smaller samples, as emphasized by our simulation study.

The second procedure is called SAMSEE (smooth AMSE estimator). This method aims to estimate the AMSE of the Hill estimator in (3.1) and its minimizer k_{opt}. To estimate the AMSE we use a preliminary estimate for the extreme value index γ and an estimator for the bias of the Hill estimator. For γ we use the generalized jackknife estimator $\hat{\gamma}_k^{GJ}$ (Gomes et al., 2001), which is asymptotically unbiased for the second order parameter $\rho = -1$. The bias estimator is constructed employing ideas from Resnick and Stărică (1997) and Danielsson et al. (2001). The latter article contrasts two extreme value estimators to approach the bias, and the former studies averaged values of $\hat{\gamma}_k$ to reduce variation within the estimates. For the bias estimator we consider the difference of two such averages,

$$\bar{b}_{\mathrm{up},K,k} := \frac{1}{K-k+1}\sum_{i=k}^{K}\hat{\gamma}_i - \frac{1}{K}\sum_{i=1}^{K}\hat{\gamma}_i.$$

Under the same conditions necessary for the asymptotic normality of the Hill estimator and if $K/k \to c > 1$, it holds that

$$\mathbb{AE}[\bar{b}_{\text{up},K,k}] = -\rho A(n/k)/(1-\rho)^2 \cdot \delta_\rho(k/K),$$

where $\delta_\rho(c) := (c^\rho - 1)/(-\rho(c^{-1} - 1)$ with $\delta_{-1}(c) = 1$. Based on this result, we construct the AMSE estimator fixing $\rho = -1$,

$$\text{SAMSEE}(k) := (\hat{\gamma}^{GJ}_{K^*})^2/k + 4(\bar{b}_{\text{up},K^*,k})^2,$$

where K^* is selected via minimizing a separate approximation error presented in article C in Section 3. The idea to fix the second order parameter $\rho = -1$ instead of including further uncertainty via estimating it is often suggested in threshold selection or bias estimation approaches and is supported by good simulation results (Gomes et al., 2001; Drees and Kaufmann, 1998; Goegebeur et al., 2008).

We compare the two new approaches to various other methods that estimate k_{opt} in a comprehensive simulation study. The numerical study illustrates the stable performance of SAMSEE across different distributions and its competitiveness with other procedures. The method IHS, which is not constructed for optimal adaptive estimation of $\hat{\gamma}_k$, still performs reasonable on this task, and it performs remarkably well in terms of quantile estimation for small samples of size $n = 500$ and up to $n = 5000$.

Finally, we apply the two methods to a time varying extreme value index $\gamma(t)$, which is a strong use case for automated data-driven threshold selection procedures. We extend the non-parametric estimator of de Haan and Zhou (2017) by including locally selected thresholds. This approach yields a simple ad hoc estimate for an extreme value index depending on a univariate covariate, and we numerically illustrate that it decreases the variation among the estimates. For further illustration, we apply it to a real-world dataset of a bank to study the severity distribution of high operational losses over time.

3.3 Discussion and Outlook

Publication C provides two new data-driven automated threshold selection procedures, which depend on no or few tuning parameters and are validated in a large comparison study. The SAMSEE approach yields an estimator for the AMSE of the Hill estimator additionally to an estimate of k_{opt}, which is a special feature of this method. Although the simulation study already accentuates the robustness and good performance of SAMSEE over all distributions, it would still be enlightening to add a consistency statement for its minimizing sequence as an estimator of k_{opt} and to derive the convergence rates of

the adaptive Hill estimates.
In article C we consider the peak-over-threshold approach, where one chooses a fixed threshold and only studies the tail above this threshold. For some features of a distribution, however, it is important to infer the tail and the bulk simultaneously. One such example is the total operational loss as proposed by the Basel Committee for Banking Supervision, which is the sum of all losses in a specific time period. Such properties can be analysed under mixture models. The basic idea is to consider different models for the bulk and the tail of the distribution and combine them, for example, by a transition function or a discontinuity at a threshold value. A review on various mixture models can be found in Scarrott and MacDonald (2012), where Bayesian approaches, which incorporate uncertainty about the threshold by letting it follow a prior distribution, are discussed as well. The prior distribution on the threshold could be an empirical prior based on SAMSEE, since it is especially useful for this task as a smooth AMSE estimator. In a similar way, in non-Bayesian models, an empirical transition function could be constructed also utilizing SAMSEE. Such an empirical transition function would make these models more flexible and data-driven.

3.4 Own Contribution

Article C is in most parts my own contribution. The construction of IHS resulted from joint discussions with the co-authors. I developed the SAMSEE procedure and did the theoretical analysis. Further, I wrote the article while benefiting from helpful comments by A. Krajina and T. Krivobokova.

Bibliography

Bader, B., Yan, J., and Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *The Annals of Applied Statistics*, 12(1):310–329.

Balkema, A. and de Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2(5):792–804.

Beirlant, J., Dierckx, G., Guillou, A., and Strărică, C. (2002). On exponential representations of log-spacings of extreme order statistics. *Extremes*, 5(2):157–180.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2004). *Statistics of Extremes – Theory and Applications*. Wiley Series in Pobability and Statistics.

Berger, J. O., Bernardo, J. M., and Sun, D. (2012). Objective priors for discrete parameter spaces. *Journal of the American Statistical Association*, 107(498):636–648.

Blumenthal, S. and Dahiya, R. C. (1981). Estimating the binomial parameter n. *Journal of the American Statistical Association*, 76(376):903–909.

Carroll, R. J. and Lombard, F. (1985). A note on n estimators for the binomial distribution. *Journal of the American Statistical Association*, 80(390):423–426.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Danielsson, J., de Haan, L., Peng, L., and de Vries, C. G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate Analysis*, 76(2):226–248.

DasGupta, A. and Rubin, H. (2005). Estimation of binomial parameters when both n, p are unknown. *Journal of Statistical Planning and Inference*, 130(1):391–404.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B*, 53(3):393–442.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory - An Introduction*. Springer.

de Haan, L. and Zhou, C. (2017). Trends in extreme value indices. working paper, https://personal.eur.nl/zhou/Research/WP/varygamma.pdf.

Dey, D. Y. and Yan, J. (2016). *Extreme Value Modeling and Risk Analysis*. Taylor and Francis Group.

Draper, N. and Guttman, I. (1971). Bayesian estimation of the binomial parameter. *Technometrics*, 13(3):667–673.

Drees, H., de Haan, L., and Resnick, S. (2000). How to make a Hill plot. *The Annals of Statistics*, 28(1):254–274.

Drees, H., Janßen, A., Resnick, S. I., and Wang, T. (2018). On a minimum distance procedure for threshold selection in tail analysis. preprint, arXiv:1811.06433.

Drees, H. and Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic processes and their Applications*, 75(2):149–172.

DuMouchel, W. H. (1983). Estimating the stable index α in order to measure tail thickness: A critique. *The Annals of Statistics*, 11(4):1019–1031.

Dupuis, D. J. (1998). Exceedances over high thresholds: a guide to threshold selection. *Extremes*, 1(3):251–261.

Ferreira, A., de Haan, L., and Peng, L. (2003). On optimising the estimation of high quantiles of a probability distribution. *Statistics*, 37(5):401–434.

Fisher, R. (1941). The negative binomial distribution. *Annals of Human Genetics*, 11(1):182–187.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190.

Gnedenko, B. V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44(3):423–453.

Goegebeur, Y., Beirlant, J., and de Wet, T. (2008). Linking Pareto-tail kernel goodness-of-fit statistics with tail index at optimal threshold and second order estimation. *REVSTAT - Statistical Journal*, 6(1):51–69.

Gomes, M. I., Martins, M. J. a., and Neves, M. (2001). Alternatives to a semi-parametric estimator of parameters of rare events - the Jackknife methodology. *Extremes*, 3(3):207–229.

Gonzalo, J. and Olmo, J. (2004). Which extreme values are really extreme? *Journal of Financial Econometrics*, 2(3):349–369.

Guillou, A. and Hall, P. (2001). A diagnostic for selecting the threshold in extreme value analysis. *Journal of the Royal Statistical Society. Series B*, 63(2):293–305.

Günel, E. and Chilko, D. (1989). Estimation of parameter *n* of the binomial distribution. *Communications in Statistics - Simulation and Computation*, 18(2):537–551.

Haldane, J. B. S. (1941). The fitting of binomial distributions. *Annals of Human Genetics*, 11(1):179–181.

Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of Multivariate Analysis*, 32(2):177–203.

Hall, P. (1994). On the erratic behavior of estimators of n in the binomial n, p distribution. *Journal of the American Statistical Association*, 89(425):344–352.

Hall, P. and Weissman, I. (1997). On the estimation of extreme tail probabilities. *The Annals of Statistics*, 25(3):1311–1326.

Hall, P. and Welsh, A. H. (1985). Adaptive estimates of the parameters of regular variation. *The Annals of Statistics*, 13(1):331–341.

Hamedani, G. G. and Walter, G. G. (1988). Bayes estimation of the binomial parameter *n*. *Communications in Statistics - Theory and Methods*, 17(6):1829–1843.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

Kahn, W. D. (1987). A cautionary note for Bayesian estimation of the binomial parameter *n*. *The American Statistician*, 41(1):38–40.

Kratz, M. and Resnick, S. I. (1996). The qq-estimator and heavy tails. *Communications in Statistics: Stochastic Models*, 12(4):699–724.

Link, W. A. (2013). A cautionary note on the discrete uniform prior for the binomial *n*. *Ecology*, 94(10):2173–2179.

Neves, C. and Fraga Alves, M. I. (2004). Reiss and Thomas' automatic selection of the number of extremes. *Computational Statistics and Data Analysis*, 47(4):689–704.

Olkin, I., Petkau, A. J., and Zidek, J. V. (1981). A comparison of *n* estimators for the binomial distribution. *Journal of the American Statistical Association*, 76(375):637–642.

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131.

Raftery, A. E. (1988). Inference for the binomial n parameter: A hierarchical Bayes approach. *Biometrika*, 75(2):223–228.

Reiss, R.-D. and Thomas, M. (2007). *Statistical Analysis of Extreme Values*. Birkhäuser Verlag.

Resnick, S. and Stărică, C. (1997). Smoothing the Hill estimator. *Advances in Applied Probability*, 29(1):271–293.

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT – Statistical Journal*, 10(1):33–60.

Schneider, L. F., Krajina, A., and Krivobokova, T. (2019). Threshold selection in univariate extreme value analysis. preprint, arXiv:1903.02517.

Schneider, L. F., Schmidt-Hieber, J., Staudt, T., Krajina, A., Aspelmeier, T., and Munk, A. (2018a). Posterior consistency for *n* in the binomial (*n*, *p*) problem with both parameters unknown - with applications to quantitative nanoscopy. preprint, arXiv:1809.02443.

Schneider, L. F., Staudt, T., and Munk, A. (2018b). Posterior consistency in the binomial (*n*, *p*) model with unknown *n* and *p*: A numerical study. preprint, arXiv:1809.02459.

Serra, P., Krivobokova, T., and Rosales, F. (2018). Adaptive non-parametric estimation of mean and autocovariance in regression with dependent errors. preprint, arXiv:1812.06948.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Villa, C. and Walker, S. G. (2014). A cautionary note on the discrete uniform prior for the binomial n: comment. *Ecology*, 95(9):2674–2677.

Addenda

The following addenda provide the three publications A, B and C that are thoroughly discussed in the previous chapters. An introductory summary is given, which lists each article's reference and abstract.

Posterior Consistency for n in the Binomial (n, p) Problem with Both Parameters Unknown - with Applications to Quantitative Nanoscopy
Laura Fee Schneider, Johannes Schmidt-Hieber, Thomas Staudt, Andrea Krajina, Timo Aspelmeier and Axel Munk
preprint available, arXiv:1809.02443, (2018)

Abstract The estimation of the population size n from k i.i.d. binomial observations with unknown success probability p is relevant to a multitude of applications and has a long history. Without additional prior information this is a notoriously difficult task when p becomes small, and the Bayesian approach becomes particularly useful.
In this paper we show posterior contraction as $k \to \infty$ in a setting where $p \to 0$ and $n \to \infty$. The result holds for a large class of priors on n which do not decay too fast. This covers several known Bayes estimators as well as a new class of estimators, which is governed by a scale parameter. We provide a comprehensive comparison of these estimators in a simulation study and extent their scope of applicability to a novel application from super-resolution cell microscopy.

Posterior Consistency in the Binomial Model with Unknown Parameters: A Numerical Study

Laura Fee Schneider, Thomas Staudt, and Axel Munk

preprint available, arXiv:1809.02459, (2018)

Abstract Estimating the parameters from k independent Bin(n, p) random variables, when both parameters n and p are unknown, is relevant to a variety of applications. It is particularly difficult if n is large and p is small. Over the past decades, several articles have proposed Bayesian approaches to estimate n in this setting, but asymptotic results could only be established recently in Schneider et al. (2018a). There, posterior contraction for n is proven in the problematic parameter regime where $n \to \infty$ and $p \to 0$ at certain rates. In this article, we study numerically how far the theoretical upper bound on n can be relaxed in simulations without losing posterior consistency.

Threshold Selection in Univariate Extreme Value Analysis

Laura Fee Schneider, Andrea Krajina, and Tatyana Krivobokova

preprint available, arXiv:1903.02517, (2019)

Abstract Threshold selection plays a key role for various aspects of statistical inference of rare events. Most classical approaches tackling this problem for heavy-tailed distributions crucially depend on tuning parameters or critical values to be chosen by the practitioner. To simplify the use of automated, data-driven threshold selection methods, we introduce two new procedures not requiring the manual choice of any parameters. The first method measures the deviation of the log-spacings from the exponential distribution and achieves good performance in simulations for estimating high quantiles. The second approach smoothly estimates the asymptotic mean square error of the Hill estimator and performs consistently well over a wide range of distributions.
The methods are compared to existing procedures in an extensive simulation study and applied to a dataset of financial losses, where the underlying extreme value index is assumed to vary over time. This application strongly emphasizes the importance of solid automated threshold selection.

CHAPTER A

Posterior Consistency for n in the Binomial (n,p) Problem with Both Parameters Unknown - with Applications to Quantitative Nanoscopy

Posterior Consistency for n in the Binomial (n, p) Problem with both Parameters Unknown - with Applications to Quantitative Nanoscopy

Laura Fee Schneider*[1], Johannes Schmidt-Hieber†[2], Thomas Staudt‡[1], Andrea Krajina§[1], Timo Aspelmeier¶[1], and Axel Munk‖[1,3]

[1]Institute for Mathematical Stochastics, University of Göttingen
[2]Mathematical Institute, Leiden University
[3]Max Planck Institute for Biophysical Chemistry, Göttingen

Abstract

The estimation of the population size n from k i.i.d. binomial observations with unknown success probability p is relevant to a multitude of applications and has a long history. Without additional prior information this is a notoriously difficult task when p becomes small, and the Bayesian approach becomes particularly useful.
In this paper we show posterior contraction as $k \to \infty$ in a setting where $p \to 0$ and $n \to \infty$. The result holds for a large class of priors on n which do not decay too fast. This covers several known Bayes estimators as well as a new class of estimators, which is governed by a scale parameter. We provide a comprehensive comparison of these estimators in a simulation study and extent their scope of applicability to a novel application from super-resolution cell microscopy.

AMS 2010 Subject Classification: Primary 62G05; secondary 62F15, 62F12, 62P10, 62P35

*laura-fee.schneider@mathematik.uni-goettingen.de
†schmidthieberaj@math.leidenuniv.nl
‡thomas.staudt@stud.uni-goettingen.de
§andrea.krajina@mathematik.uni-goettingen.de
¶timo.aspelmeier@mathematik.uni-goettingen.de
‖munk@math.uni-goettingen.de

Keywords: Bayesian estimation, posterior contraction, binomial distribution, beta-binomial likelihood, quantitative cell imaging, improper prior

1 Introduction and motivation

Presumably, the binomial distribution $\mathrm{Bin}(n, p)$ is the most fundamental and simple model for the repetition of independent success/failure events. When both parameters p and n are unknown, which is the topic of this paper, it serves as a basic model for many applications. For example, n corresponds to the population size of a certain species (Otis et al., 1978; Royle, 2004; Raftery, 1988), the number of defective appliances (Draper and Guttman, 1971) or the number of faults in software reliability (Basu and Ebrahimi, 2001). In Section 4 we elaborate on a novel application where n is the number of unknown fluorescent markers in quantitative super-resolution microscopy (Hell, 2009; Aspelmeier et al., 2015).

Accordingly, joint estimation of the population size n and the success probability p of a binomial distribution from k independent observations has a long history dating back to Fisher (1941). In contrast to the problem of estimating p or n when one of the parameters is known (Lehmann and Casella, 1996), this is a much more difficult issue. Fisher suggested the use of the sample maximum (which is a consistent estimator for n as $k \to \infty$) and argued that the estimator is always "good", as long as the sample size is large enough. In fact, if $X_1, \dots, X_k \overset{i.i.d.}{\sim} \mathrm{Bin}(n, p)$ for fixed n and p, the sample maximum converges exponentially fast to n as $k \to \infty$ since

$$\mathbb{P}\big(\max_{i=1,\dots,k} X_i = n\big) = 1 - \mathbb{P}\big(\max_{i=1,\dots,k} X_i < n\big) = 1 - (1 - p^n)^k. \tag{1.1}$$

While true asymptotically, the maximum very strongly underestimates the true n even for relatively large sample size k if the probability of success is small. This is explicitly quantified in DasGupta and Rubin (2005): if $p = 0.1$ and $n = 10$, then the sample size k needs to be larger than 3635 to ensure that $\mathbb{P}(\max_{i=1,\dots,k} X_i \geq n/2) \geq 1/2$. If $p = 0.1$ and $n = 20$, one would need a sample size of more than $k = 900{,}000$ to guarantee the same probability statement as above.

This fallacy of the sample maximum can be explicitly seen in a refined asymptotic analysis for n and p as well. By Bernoulli inequality and since $1 - x \leq e^{-x}$, it follows from (1.1) that

$$1 - e^{-kp^n} \leq \mathbb{P}\big(\max_{i=1,\dots,k} X_i = n\big) \leq kp^n,$$

which means that if $kp^n \to 0$, the sample maximum is no longer a consistent estimator of n. This occurs, for example, in the domain of attraction of the Poisson distribution, i.e.,

when $n \to \infty,\ p \to 0$ and $np \to \mu \in (0, \infty)$ as $k \to \infty$ and $\log(k) \leq n$, since

$$\begin{aligned} kp^n &= \exp\{\log(k) + n\log(p)\} \sim \exp\{\log(k) - n\log(n)\} \\ &\leq \exp\{\log(k) - \log(k)\log(\log(k))\} \to 0, \text{ as } k \to \infty. \end{aligned}$$

In fact, when $np \to \mu$ both parameters become indistinguishable and this asymptotic scenario serves as a limiting benchmark for the $\text{Bin}(n, p)$ problem to become solvable. However, in many applications the small p regime (rare events) is the relevant one (see the references below and Section 4), and this will be the topic of this paper.

A variety of methods addressing this issue and improving over the sample maximum have been provided over the last decades but a final answer remains elusive until today. Broadly speaking, a major lesson from these attempts to obtain better estimators (see Section 2.1 for a detailed discussion) seems that in this difficult regime further information on n and p is required to obtain estimators performing reasonably well. This asks for a Bayesian approach. An early Bayesian estimator of the binomial parameters (N, P), now considered as random, dates back to Draper and Guttman (1971), who suggested the mode of the posterior distribution for a uniform prior on $\{1, \ldots, N_0\}$ for N and a $\text{Beta}(a, b)$ prior for P. Here, $N_0 \in \mathbb{N}$ is fixed and the parameters $a, b > 0$ are usually chosen as $a = b = 1$, which yields the standard uniform distribution. Later Raftery (1988), Günel and Chilko (1989), Hamedani and Walter (1988) and Berger et al. (2012) provided further estimators, which mainly differ in their choices of loss functions and prior distributions for N and P. A hierarchical Bayes approach is introduced in Raftery (1988) with a Poisson prior on N with mean μ, which implies a Poisson distribution with parameter $\lambda = \mu p$ as the marginal distribution of each observation. The prior for the pair (λ, P) is chosen proportional to $1/\lambda$, which is equivalent to a product prior for the pair (N, P) with the prior for N proportional to $1/n$ and the standard uniform prior for P. Raftery (1988) suggested to minimize the Bayes risk with respect to the relative quadratic loss, which seems particularly suitable for estimating n and will be adopted in this paper as well. From extensive simulation studies (see the afore mentioned references and Section 3), it is known that such Bayesian estimators of n deliver numerically good results, in general. However, to the best of our knowledge, there is no rigorous theoretical underpinning of these findings. In particular, nothing is known about the posterior concentration of such estimators, and no systematic understanding of the role of the prior has been established.

Our contribution to this topic is threefold: (i) we propose a new class of Bayesian estimators for n, generalizing the approach in Raftery (1988), and (ii) we prove the posterior contraction for n. The posterior contraction result holds for a wide class of priors for n and does not depend on the choice of the loss function. It implies consistency in a general asymptotic setting of the introduced class of estimators as well as of many (and with small

changes even all) Bayesian estimators mentioned above. Finally (iii), we extend the i.i.d. $\mathrm{Bin}(n,p)$ model to a regression setting and apply our Bayes approach to count the number of fluorophores from super-resolution images, which is considered a difficult task.

Ad (i). For the new class of estimators, which we call the *scale estimators*, we consider k independent random variables $X_1,\dots,X_k$ from a $\mathrm{Bin}(N,P)$ distribution. Denote $\mathbf{X}^k := (X_1,\dots,X_k)$ and $M_k := \max_{i=1,\dots,k} X_i$. We assume a product prior for the pair (N,P), where the prior for P is $\Pi_P \sim \mathrm{Beta}(a,b)$ for some $a,b>0$, and Π_N, the prior for N, satisfies $\Pi_N(n) \propto n^{-\gamma}$ for $\gamma > 1$. Independence of N and P is a common assumption and also justified in our example (Section 4) based on physical considerations. The scale estimator is then defined as the minimizer of the Bayes risk with respect to the relative quadratic loss, $l(x,y) = (x/y-1)^2$. Following Raftery (1988), it is given by

$$\hat{n} := \frac{\mathbb{E}\left[\frac{1}{N}|\mathbf{X}^k\right]}{\mathbb{E}\left[\frac{1}{N^2}|\mathbf{X}^k\right]} = \frac{\sum_{n=M_k}^{\infty} \frac{1}{n} L_{a,b}(n)\Pi_N(n)}{\sum_{n=M_k}^{\infty} \frac{1}{n^2} L_{a,b}(n)\Pi_N(n)}, \tag{1.2}$$

where $L_{a,b}(n)$ is the beta-binomial likelihood, see, e.g., Carroll and Lombard (1985). In existing literature (Berger et al., 2012; Link, 2013), the Bayesian estimator of n with the prior $\Pi_N(n) \propto 1/n$ is often called the scale estimator. Even though we do not allow $\gamma = 1$ in the above definition since it leads to an improper prior (see, however, Theorem 2 for a proper modification of these estimators for $0 \le \gamma \le 1$, which makes them accessible to our theory), we adopt this name for the new class of estimators.

Ad (ii). We show posterior consistency in a quite general setting, where the prior distribution Π_N can be chosen freely as long as it is a well-defined probability distribution satisfying

$$\Pi_N(n) \ge \beta e^{-\alpha n^2},\ n \in \mathbb{N} \tag{1.3}$$

for some positive constants α and β. In our asymptotic setting we consider sequences of parameters $(n_k,p_k)_k$ that may depend on the sample size k and are described by the class

$$\mathcal{M}_\lambda := \left\{(n_k,p_k)_k : 1/\lambda \le n_k p_k \le \lambda,\ n_k \le \lambda \sqrt[6]{k/\log(k)}\right\}, \tag{1.4}$$

for $\lambda > 1$. We show that

$$\sup_{(n_k^0,p_k^0)\in\mathcal{M}_\lambda} \mathbb{E}_{n_k^0,p_k^0}\left[\Pi\big(N \ne n_k^0 \mid \mathbf{X}^k\big)\right] \to 0,\ \text{as } k\to\infty,$$

where $X_1,\dots,X_k \overset{i.i.d.}{\sim} \mathrm{Bin}(n_k^0,p_k^0)$. This is the main result of the paper and it will be formally stated as Theorem 1 in Section 2.

The recent advances on posterior contraction focus mainly on nonparametric or semiparametric models (Ghosal et al., 2000; Ghosal and van der Vaart, 2017) and posterior contraction for model selection in high-dimensional setups (Castillo and van der Vaart, 2012; Castillo et al., 2015; Gao et al., 2015). Discrete models with complex structure have not yet been studied and it appears difficult to approach them by a general treatment. Our proof uses earlier work on maximum likelihood estimation by Hall (1994) and opens another route to establish posterior consistency beyond the standard approach via testing, see Schwartz (1965).

In the binomial model, posterior consistency for fixed parameters n and p with the priors above follows already by Doob's consistency theorem, see, e.g., van der Vaart (1998). To the best of our knowledge, no refined asymptotic result for a Bayesian approach to estimate n when p is unknown exists. Our result shows consistency of the marginal posterior distribution of N even in the challenging and relevant case of $n_k \to \infty$ and $p_k \to 0$ as $k \to \infty$ as long as $(n_k, p_k)_k \in \mathcal{M}_\lambda$. The difficulty of this setup comes from the convergence of the binomial distribution to the Poisson distribution with parameter $\mu = \lim_{k\to\infty} np$ as $n = n_k \to \infty$ and $p = p_k \to 0$. We have seen that the sample maximum is consistent as long as $kp^n \to \infty$, for which $e^n = o(k)$ is necessary (but not sufficient, see Lemma 5 for more details). In contrast, the definition of the class $\mathcal{M}_\lambda$ implies that $n^{6+\epsilon} = O(k)$ for $\epsilon > 0$ is already sufficient for the posterior consistency of the suggested Bayes approach. We stress that a simulation study in Schneider et al. (2018) suggests that the rate in Theorem 1 cannot be relaxed significantly, as numerically posterior consistency is only observed up to $n^4 = O(k)$.

The posterior contraction result holds for the introduced scale estimators with $\Pi_N(n) \propto n^{-\gamma}$, $\gamma > 1$. The improper priors with $0 \leq \gamma \leq 1$ satisfy the assumptions under slight modifications, which are described in Theorem 2 in Section 2. With these modifications (restricting the support of N) the estimators of Draper and Guttman (1971) and Raftery (1988) are also covered by our theory. Our Theorems are applicable to many other Bayes estimators, as well. For example, Theorem 1 holds for the estimator in Günel and Chilko (1989), where a Gamma prior for N is suggested, and for the estimator in Hamedani and Walter (1988), which suggests either a poisson prior on N or an improper prior that can be considered via Theorem 2.

Ad (iii). Modern cell microscopy allows visualizing proteins and their modes of interaction during activity. It has become an indispensable tool for understanding biological function, transport and communication in the cell and its compartments, especially since the development of super-resolution nanoscopy (highlighted by the 2014 Nobel Prize in

Chemistry). These techniques enable imaging of individual proteins through photon counts obtained from flourescent markers (fluorophores), which are tagged to the specific protein of interest and excited by a laser beam (see Hell (2015) for a recent survey). In this paper, we are concerned with single marker switching (SMS) microscopy (Betzig et al., 2006; Rust et al., 2006; Hess et al., 2006; Fölling et al., 2008) where the emission of photons, which are then recorded, is inherently random: after laser excitation a fluorophore undergoes a complicated cycling through (typically unknown) quantum mechanical states on different time scales. This severely hinders a precise determination of the number of molecules at a certain spot in the specimen, see, e.g., Lee et al. (2012), Rollins et al. (2015), Aspelmeier et al. (2015). In Section 4 we show how the number of fluorophores can be obtained from a modified (n, p)-Binomial model when they occur in clusters of similar size in the biological sample. A common difficulty in such experiments is that the number of active markers decreases over the measurement process due to bleaching effects. We show that the initial number $n^{(0)}$ can still be estimated from observations $X^{(t)} \sim \mathrm{Bin}(n^{(t)}, p)$ at different time points t. We can link $n^{(0)}$ to $X^{(t)}$ by an exponential decay $n^{(t)} = n^{(0)}(1 - B)^t$, which is known to be valid on physical grounds. This results now in a variant of the (n, p)-Binomial model, where the bleaching probability B of a fluorophore can be estimated jointly with $n^{(0)}$ within this model. This allows us to determine the number of fluorophores $n^{(0)}$ on DNA origami test beds with high accuracy.

This paper is organized as follows. Our main result on posterior contraction and the discussion on the asymptotics of other estimators for n can be found in Section 2. Section 3 contains an extensive simulation study comparing the finite sample properties of those estimators and investigating robustness against model deviations from the $\mathrm{Bin}(n, p)$ model relevant to our data example. In Section 4 the data example is presented. The proof of the posterior contraction and some auxiliary results about binomial random variables are stated in Section 5. Further auxiliary technicalities are deferred to the Appendix A.

2 Posterior contraction for n

Throughout the following $X_1, \ldots, X_k$ are independent random variables with a $\mathrm{Bin}(N, P)$ distribution. We assume a product prior $\Pi_{(N,P)} = \Pi_N \Pi_P$ for the pair (N,P). For P we choose a $\mathrm{Beta}(a, b)$ prior with parameters $a, b > 0$. It is the conjugate prior suggested in Draper and Guttman (1971) and widely used. The prior Π_N for N can be chosen as any proper probability distribution on the positive integers such that (1.3) holds for some $\alpha, \beta > 0$. Write $\mathbf{X}^k = (X_1, \ldots, X_k)$, $M_k = \max_{i=1,\ldots,k} X_i$ and $S_k := \sum_{i=1}^k X_i$. For $A \subset [0, 1]$

and $n \in \mathbb{N}$, the joint posterior distribution for P and N is then given by

$$\Pi\big(P \in A, N = n \,|\, \mathbf{X}^k\big) = \frac{\int_A t^{S_k+a-1}(1-t)^{kn-S_k+b-1}\mathrm{d}t \cdot \prod_{i=1}^k \binom{n}{X_i} \cdot \Pi_N(n)}{\sum_{m=1}^\infty \int_0^1 t^{S_k+a-1}(1-t)^{km-S_k+b-1}\mathrm{d}t \cdot \prod_{i=1}^k \binom{m}{X_i} \cdot \Pi_N(m)}$$

if $n \geq M_k$ and $\Pi(P \in A, N = n \,|\, \mathbf{X}^k) = 0$ otherwise. The marginal posterior likelihood function for N is thus

$$\Pi\big(N = n \,|\, \mathbf{X}^k\big) \propto \prod_{i=1}^k \binom{n}{X_i} \frac{\Gamma(kn - S_k + b)\Gamma(S_k + a)}{\Gamma(kn + a + b)} \mathbf{1}(n \geq M_k)\Pi_N(n) =: L_{a,b}(n)\Pi_N(n),$$

where $\mathbf{1}(\cdot)$ denotes the indicator function and $L_{a,b}(\cdot)$ is the beta-binomial likelihood, see, e.g., Carroll and Lombard (1985).

The main result is stated in the following theorem and shows posterior contraction for n in the asymptotic setting described by sequences of parameters $(n_k, p_k)_k \in \mathcal{M}_\lambda$ as defined in equation (1.4).

Theorem 1. *Conditionally on $N = n_k^0$ and $P = p_k^0$ let $X_1, \ldots, X_k \overset{i.i.d.}{\sim} \mathrm{Bin}(n_k^0, p_k^0)$. For any prior distribution $\Pi_{(N,P)} = \Pi_N \Pi_P$ with $\Pi_P = \mathrm{Beta}(a, b)$, $a, b > 0$, and where Π_N is a probability distribution such that (1.3) holds, we have uniform posterior contraction over $\mathcal{M}_\lambda$ in (1.4) for $\lambda > 1$, i.e.,*

$$\sup_{(n_k^0, p_k^0)_k \in \mathcal{M}_\lambda} \mathbb{E}_{n_k^0, p_k^0}\big[\Pi\big(N \neq n_k^0 \,|\, \mathbf{X}^k\big)\big] \to 0, \text{ as } k \to \infty.$$

As mentioned in the introduction, from Theorem 1 follows posterior consistency for the Bayesian estimators in equation (1.2) and the ones in Hamedani and Walter (1988) and Günel and Chilko (1989), when considering parameter sequences in $\mathcal{M}_\lambda$. The estimator in Draper and Guttman (1971) is based on a beta prior for P and a uniform prior on $\{1, \ldots, N_0\}$ for some $N_0 \in \mathbb{N}$ for N. Since $n_k > N_0$ cannot be excluded for k large enough, assumption (1.3) is not fulfilled in this case. The estimators in Raftery (1988), Berger et al. (2012) and Link (2013) violate the conditions of Theorem 1 as well, since they are based on an improper prior on N proportional to $1/n$. However, we can still extend our result to modifications of these estimators, where the support of N is bounded but increases with k.

Theorem 2. *Theorem 1 holds if we exchange Π_N by $\Pi_{N,k}(n) \propto \frac{1}{n^\gamma}\mathbf{1}_{[1,T_k]}(n)$ with $\gamma \in [0, 1]$, where T_k satisfies*

$$\lambda \sqrt[6]{k/\log(k)} \leq T_k < \begin{cases} \left(\exp\left\{\alpha k^{1/3}\right\}/\beta\right)^{\frac{1}{1-\gamma}}, & \gamma < 1, \\ \exp\left\{\exp\left\{\alpha k^{1/3}\right\}/\beta\right\}, & \gamma = 1, \end{cases} \tag{2.1}$$

for all k and some positive constants α and β.

Remark 1. *Theorem 1 and Theorem 2 still hold true if we allow λ in $\mathcal{M}_\lambda$ to increase with k, as long as $\lambda_k = o\left(\log(k)^{1/14}\right)$. This statement follows by verifying the conditions on the constants in the proof of Theorem 1 and their dependence on λ. The strongest restriction results from equation (5.8) and depends on Lemma 3.*

2.1 Asymptotic results for frequentist methods: constrasted and compared

In the following we present various existing asymptotic results for frequentist estimators and put them into perspective to Theorem 1, highlighting the differences of their respective asymptotic settings to the one described by the set $\mathcal{M}_\lambda$.

Early estimators based on the method of moments and the maximum likelihood approach can be found in Haldane (1941) and Blumenthal and Dahiya (1981). Their properties are further studied in Olkin et al. (1980), where it is shown that the estimators for n, when both n and p are unknown and p is small, are highly irregular and stabilized versions of the two estimators are proposed. Two estimators were introduced more recently in DasGupta and Rubin (2005): the first one is another modification of the method of moments estimator, and the second one is a bias correction of the sample maximum. The asymptotic behavior of these two estimators is also known. For the new moments estimator, $\hat{n}_{\mathrm{NME}}$, it holds that, as $k \to \infty$,

$$\sqrt{k}(\hat{n}_{\mathrm{NME}} - n) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2\gamma^2 n(n-1)),$$

where n is fixed and $\gamma > 0$ is a tuning parameter to be chosen by the practitioner. For the bias corrected sample maximum, $\hat{n}_{\mathrm{bias}}$ say, it holds for n fixed, as $k \to \infty$:

$$(nk)^{1/(n-1)}(\hat{n}_{\mathrm{bias}} - n) \xrightarrow{\mathcal{D}} \delta_1,$$

where δ_1 denotes the Dirac measure at 1.

In Carroll and Lombard (1985) a further modification of the maximum likelihood estimator is introduced. The estimator is the maximizer of the beta-binomial likelihood for n, where a beta density is assumed for p and p is integrated out. The Carroll-Lombard estimator is nearly equivalent to the Bayesian Draper-Guttman estimator (e.g., for N_0 large they produce the same estimates), since the Carroll-Lombard estimator can be understood as a maximum a posteriori (MAP) Bayesian estimator of n with an improper uniform prior on $\mathbb{N}$. That means, if we bound the set of values where a maximum can be attained to $\{1, \ldots, T_k\}$, then Theorem 2 with $\gamma = 0$ applies to the Caroll-Lombard estimator as well. This extends the classical asymptotic normality result of the Carroll-Lombard estimator

$\hat{n}_{\mathrm{CL}}$, which holds for p constant, $n \to \infty$ and $\sqrt{k}/n \to 0$ as $k \to \infty$:

$$\sqrt{k}\left(\frac{\hat{n}_{\mathrm{CL}} - n}{n}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, 2(1-p)^2/p^2\big).$$

All of the results above hold for either n or p fixed and hence provide only limited insight into the situation when p is small. A notable extension is discussed in Hall (1994). There, it is shown for $n = n_k \to \infty$ and $p = p_k \to 0$ as $k \to \infty$ that

$$\frac{p\sqrt{k}}{2}\left(\frac{\hat{n}_{\mathrm{CL}} - n}{n}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

if $np \to \mu \in (0, \infty)$ and $kp^2 \to \infty$ as $k \to \infty$. Note that this result, like the previous, studies the limiting distribution of the relative difference, where $\hat{n} - n$ is scaled by n. In contrast, we show posterior contraction to the exact value $\hat{n} = n$. This explains that Hall (1994) can allow a faster rate ($n = o(k^2)$) than in our setting where $n = O\big(\sqrt[6]{k/\log(k)}\big)$. Also note that the above result is one specific scenario in a broader context and relies on further technical conditions, like n to be lower bounded by some positive power of k.

3 Simulation study

In this section we investigate the finite sample performance of Bayesian estimators numerically for different choices of priors Π_P and Π_N. We compare the following estimators that we introduced in the previous sections.

(SE) The scale estimator SE(γ) with $\Pi_P = \mathrm{Beta}(a, b)$ and $\Pi_N(n) \propto n^{-\gamma}$. We consider both proper prior distributions ($\gamma > 1$), and improper ones ($0 \leq \gamma \leq 1$). The beta prior is chosen such that P has expectation $\hat{p}$, where $\hat{p} \in (0, 1]$ is a first guess for the probability of success, which might roughly be known beforehand. We select a to be 1 or 2 and set $b = b(\hat{p}) := a/\hat{p} - a$. The scale factor γ needs to be chosen. Note that the Raftery estimator is equivalent to the scale estimator with $\gamma = 1$ and $a = b = 1$.

(DGE) The Draper-Guttman estimator DGE(N_0). The parameters a and b of the beta distribution are selected in the same way as for the scale estimator. The upper bound N_0 should be selected sufficiently large to avoid underestimation.

We look at the SE with $\gamma \in \{0, 0.5, 1, 2, 3\}$ and the DGE with $N_0 = 500$. In case of an improper prior ($\gamma \leq 1$), Theorem 2 applies, and the posterior distribution is well defined as long as $a + \gamma > 1$ (see Kahn (1987) for a cautionary note on this problem). We also employ the estimator SE(0) with $a = 1$, for which the posterior does not exist (so it is no Bayes estimator), but which still produces finite estimates.

General performance. Our first simulation study is based on 1000 samples of size $k \in \{30, 100, 300\}$ from a binomial distribution $\text{Bin}(n_0, p_0)$ for $n_0 \in \{20, 50\}$ and $p_0 \in \{0.05, 0.1, 0.3\}$. For all pairs (n_0, p_0) and each estimator $\hat{n}$ we simulate:

- the relative mean squared error (RMSE), given by $\mathbb{E}\left[\left(\frac{\hat{n}}{n_0} - 1\right)^2\right]$,
- the bias $\mathbb{E}[\hat{n}] - n_0$ of the estimator.

We set $\hat{p} = p_0$ in the beta prior for this simulation and study the influence of the parameter γ. In Table 1, we present the estimators that have the lowest RMSE and the lowest bias for the different choices of k. The outcome advises to select smaller values of γ, the smaller p_0 is expected to be. Note that the DG estimator with large values N_0 is similar to the MAP estimator with the improper prior $\gamma = 0$. Thus, it is not surprising that there is only little difference between the performance of DG(500) and SE(0) in the simulations. Both of them perform superior in the regime of very small p_0. Still, one should be aware that a small γ increases the variance of the posterior and therefore of the estimates. For this reason, higher choices of γ become preferable for low RMSEs as k increases. The similarity of Table 1 (A) and (B) for $n_0 = 20$ and $n_0 = 50$ suggests that the influence of n_0 is much weaker than the one of p_0 for the optimal estimator choice.

(A) $n_0 = 20$

p_0	k	RMSE	bias
0.05	30	DGE(500)	SE(0)
0.05	100	DGE(500)	SE(0)
0.05	300	SE(0.5)	DGE(500)
0.1	30	DGE(500)	SE(0)
0.1	100	SE(0.5)	DGE(500)
0.1	300	SE(1)	DGE(500)
0.3	30	SE(2)	SE(1)
0.3	100	SE(3)	DGE(500)
0.3	300	SE(3)	SE(2)

(B) $n_0 = 50$

p_0	k	RMSE	bias
0.05	30	DGE(500)	SE(0)
0.05	100	DGE(500)	SE(0)
0.05	300	SE(0.5)	DGE(500)
0.1	30	DGE(500)	SE(0)
0.1	100	SE(0.5)	DGE(500)
0.1	300	SE(1)	SE(0.5)
0.3	30	SE(1)	SE(0.5)
0.3	100	SE(3)	DGE(500)
0.3	300	SE(3)	DGE(500)

TABLE 1: Overview of the estimators with the smallest RMSE and the smallest absolute bias for $a = 2$ and $b = 2/p_0 - 2$.

Our next numerical study covers a setting that is motivated by the data example in Section 4, where $p_0 \approx 0.0339$ and $k = 94$. We therefore set $p_0 = 0.0339$ and $k = 94$, and we select $n_0 = 15$. Our focus lies on the effect of the parameters a and b, and particularly

on the stability of the results with respect to misspecification of the guess $\hat{p}$. To this end, we consider four different scenarios: no information about p_0 (setting $\hat{p} = 0.5$), perfect information ($\hat{p} = p_0$), underestimation ($\hat{p} = 0.5\,p_0$), and overestimation ($\hat{p} = 1.5\,p_0$).

The results in Table 2 show that it is advantageous to choose a small γ and a unimodal beta prior (i.e., $a = 2$) if p_0 is known. If we have no information or are overestimating, it is again advisable to select $\gamma = 0$, while choosing a less confident prior for P with $a = 1$. In contrast, underestimation of p_0 leads to high instabilities and substantial overestimation of n_0 if γ is small. Here, estimators with proper priors for $\gamma = 1$ and 2 perform very well: the tendency for overestimation caused by the choice $\hat{p} = 0.5\,p_0$ is compensated by the tendency for underestimation in case of higher values of γ.

$\hat{p}$	a	estimator	RMSE	bias
0.5	1	SE(0.5)	0.478	-10.17
	1	SE(0)	0.395	-9
p_0	2	DGE(500)	0.034	-0.266
	2	SE(0)	0.036	-0.043

$\hat{p}$	a	estimator	RMSE	bias
$1.5\,p_0$	1	SE(0)	0.12	-3.73
	2	SE(0)	0.121	-4.69
$0.5\,p_0$	1	SE(1)	0.036	-0.032
	2	SE(2)	0.025	-0.55

TABLE 2: The two estimators that perform best under different choices of $\hat{p}$ for $n_0 = 15$, $p_0 = 0.0339$, and $k = 94$. The respective values of b are given by $b(\hat{p}) = a/\hat{p} - a$.

The general lesson seems to be that the smaller p_0, the more difficult it becomes to estimate n_0 and the smaller we want to choose γ. A smaller γ, however, increases the variance of the posterior distribution and leads to estimators that are more sensitive against misspecification of $\hat{p}$ in the beta prior. This is investigated in Table 3, where we compare the sensitivity of estimators corresponding to $\gamma = 0$ and $\gamma = 1$. We see that misspecifying $\hat{p} = 0.5\,p_0$ leads to severe overestimates $\mathbb{E}[\hat{n}] \approx 2\,n_0$ for DGE(500), while SE(1) is less sensitive. Selecting $\gamma = 0$ can therefore help to estimate n_0 in very difficult scenarios, but it can also lead to heavily biased results if $\hat{p}$ is chosen too small.

Robustness. Motivated by our data example in Section 4, we also investigate the situation where n may slightly vary within the sample. This appears to be relevant in many other situations as well, e.g., the (unknown) population size of a species may vary from experiment to experiment in the capture-recapture method. Whereas varying probabilities p have been investigated in Basu and Ebrahimi (2001), models with a varying population size n have not received any attention in the previous research.

We consider 1000 repetitions of size $k = 100$, where each observation X_i, $i = 1, \ldots, k$, is generated from a $\text{Bin}(n_i, p_0)$ distribution. Each n_i is in turn a realization of a binomial

estimator	$\hat{p}$	RMSE	bias
SE(1)	p_0	0.122	-4.85
	$0.5\,p_0$	0.129	4.43
	$1.5\,p_0$	0.279	-7.73
DGE(500)	p_0	0.034	-0.27
	$0.5\,p_0$	1.002	14.32
	$1.5\,p_0$	0.139	-5.09

TABLE 3: Sensitivity of SE(1) and DGE(500) against misspecification of $\hat{p}$. The value a is set to 2. All other parameters are selected like in Table 2. Note that the behavior of DGE(500) and SE(0) is comparable in this setting.

random variable $N \sim \text{Bin}(\tilde{n}, \tilde{p})$. For each sample, p_0 is drawn from a $\text{Beta}(2, 38)$ distribution with expectation 0.05. To test the influence of the varying parameter n_i, we compare the performance of the estimators in the described scenario to their performance on binomial samples with a constant n_0 (chosen as the integer nearest to $\mathbb{E}[N] = \tilde{n}\tilde{p}$) and the same realizations p_0. We calculate the RMSE with respect to n_0 for both scenarios and present the RMSE for $X_i \sim \text{Bin}(n_i, p)$ divided by the RMSE in the i.i.d. case. The ratios in Table 4 verify a stable performance of the estimators in this setting since all values are close to 1. The parameters in Table 4 are chosen close to the data example in Section 4 with $\tilde{n} \in \{8, 22\}$ and $\tilde{p} = 0.7$, but further simulations (not shown) confirmed the stability for other parameter choices, like $\tilde{p} = 0.5$ or $\tilde{p} = 0.9$, as well. Hence, in summary, we find that for inhomogeneous (random) N all estimators perform quite similar to the situation of a homogeneous (constant) n_0 ($\approx \mathbb{E}[N]$).

	$\tilde{n} = 8$	$\tilde{n} = 22$
estimator	RMSE-R	RMSE-R
SE(0.5)	1.022	1.130
SE(1)	1.011	1.067
SE(2)	1.020	1.010
DGE(500)	1.032	1.073
RE	0.988	0.981

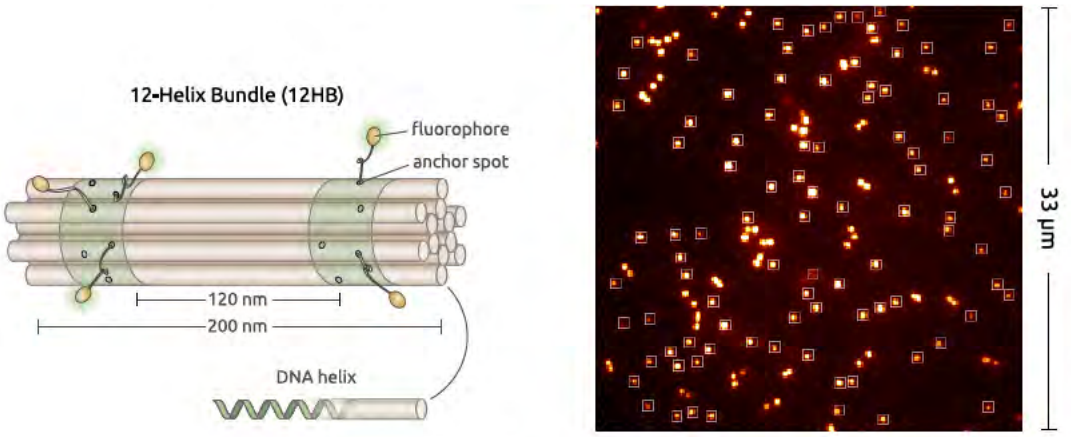
TABLE 4: Ratios of the RMSE for i.i.d. and non-i.i.d. samples (RMSE-R) for the estimators SE(γ), DGE(N_0), and the Raftery estimator RE. The beta prior in SE and DGE is defined by $a = 2$ and $b = 38$.

4 Data example

In this section we extend the previously described Bayesian estimation methods to quantify the number of fluorescent molecules in a specimen recorded with super-resolution microscopy. Reliable methods to count such molecules are highly relevant to quantitative cell biology, for example, to determine the number of proteins of interest in a compartment of the cell, see, e.g., Lee et al. (2012), Rollins et al. (2015), Ta et al. (2015), Aspelmeier et al. (2015) or Karathanasis et al. (2017) and references therein.

Experimental setup. Data has been recorded at the Laser-Laboratorium Göttingen e.V. During experimental preparation so called DNA origami (Schmied et al., 2014), tagged with the fluorescent marker Alexa647, were dispersed on a cover slip. DNA origami are nucleotide sequences which have been engineered in such a way that the origami folds itself into a desired shape (see Fig. 1A). Fluorescent molecules (fluorophores), which are equipped with an "anchor" that sticks to a specific region of the origami, are attached to the origami molecules. In the experiment, Alexa647 fluorophores with 22 different types of anchors were used, each one matching a different anchor position on the origami (see Fig. 1A). Therefore, at most 22 fluorophores can be attached to a single origami. The pairing itself is random (so not every possible anchor position needs to be occupied) and is expected to occur with a probability between 0.6 and 0.75, according to producer specifications.

Fundamental to super-resolution microscopy is the switching behavior of the fluorophores. A fluorophore can be in two different states ("on" or "off") but only emits light in the "on" state. When excited with a laser beam, it switches between these "on" and "off" states until it bleaches, i.e., reaches an irreversible "off" state. During the course of the experiment, an image sequence of several origami distributed on a cover slip is recorded over a period of a few minutes (see the movie supplement material). The exposure time for one image (denoted as frame) is 15 ms. Switching of fluorophores between "on" and "off" states is necessary to achieve super-resolution, which denotes the ability to discern markers with distance below the diffraction limit achievable with visible light of about 250 – 500 nm (Hell, 2009). Such fluorophores could not be discerned by conventional microscopy. Super-resolution becomes possible by separating photon emissions of spatially close molecules in time. This is realized by applying a low laser intensity, such that only a small fraction of fluorophores switches in the "on" state for a given frame. Hence, it is very unlikely that nearby fluorophores emit photons at the same time (see, e.g., Betzig et al. (2006), Rust et al. (2006), Hess et al. (2006), Fölling et al. (2008) for different variants of this principle). By this method, an increased resolution of up to 20 – 30 nm can be achieved.



(A) (B)

FIGURE 1: (A) Schematic drawing of the DNA-origami used in the experiment. The origami is a tube-like structure that consists of 12 suitably folded DNA helices. In each of the two highlighted green regions up to 11 fluorescence markers can anchor. (B) First frame from the sequence of microscopic images. The 94 regions of interest (ROIs) that were chosen for analysis are identified by white boxes. The selection was done algorithmically. No overlapp between ROIs was allowed, and it was made sure that no excessive background noise and disturbances affected the ROI during the course of the experiment.

The experiment was prepared in such a way that most fluorophores are guaranteed to be "on" in the first frame, and all origami are thus visible as bright spots in Fig. 1B. Note that individual fluorophores occupying the same origami cannot be discerned in this frame. This becomes possible only when most of the fluorophores are switched "off" at later times, such that markers show up individually (see the supplementary movie for illustration).

Quantitative biology. Quantitative biology addresses the issue of counting the number of fluorophores from measurements like the one described above. The brightness of each spot is proportional to the number of fluorophores in the "on" state within the respective origami. An origami is invisible if all of its fluorophores are "off", but its location is still known from the first frame, which allows us to register 94 regions of interest (ROIs) in a preparational step (see Figure 1B). Exemplarily, six microscopic frames (out of 14,060) recorded at different times $t \in \{1500, 3000, 4500, 6000, 7500, 9000\}$, which show the influence of switching and bleaching on the observations, are visualized in Figure 2.

We aim to estimate the number of fluorophores attached to each origami, which we expect to be between 13 and 16 according to the producer specification. In order to make our model

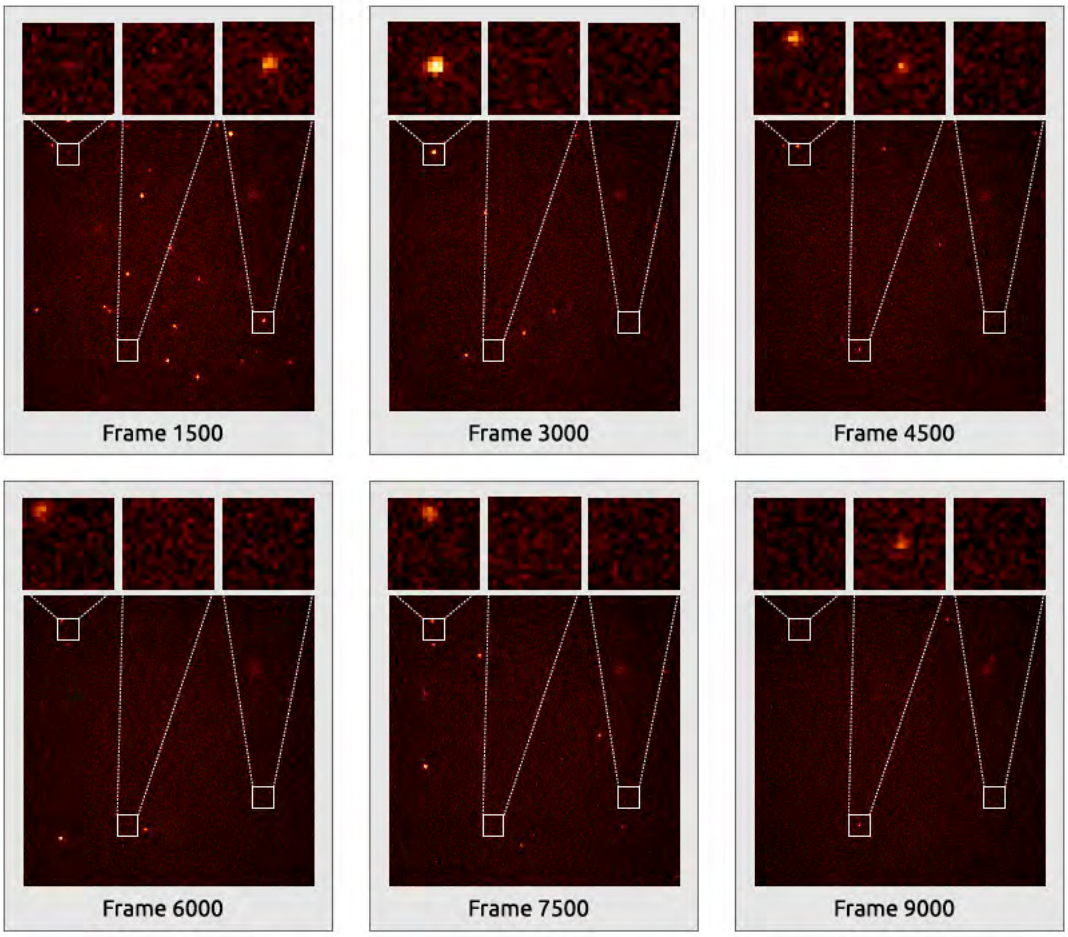


FIGURE 2: Six selected frames from the dataset of recorded origami. The (physical) time difference between two consecutive images in this figure is roughly 22.5 seconds. Bleaching causes the number of visible origami to decrease with increasing frame number, and switching causes that unbleached origami are visible only in some frames.

accessible to the data, we assume for simplicity that each origami carries the same number $n^{(0)}$ of fluorophores and we only model the mean number $n^{(t)}$ of unbleached fluorophores at time t. The physical relation between $n^{(0)}$ and $n^{(t)}$ is given by

$$n^{(t)} = n^{(0)}(1-B)^t, \tag{4.1}$$

where B denotes the bleaching probability. Now, the brightness observed for a spot in frame t is proportional to the number $X^{(t)}$ of "on" fluorophores during the frame's exposure. This number $X^{(t)}$ is binomially distributed $\mathrm{Bin}(n^{(t)}, p)$, where p denotes the (time-independent) probability that an unbleached fluorophore is in its "on" state. We can estimate $n^{(0)}$ and B by fitting a log-linear model to equation (4.1), where the respective population sizes $n^{(t)}$

are in turn estimated from the 94 realizations of $X^{(t)}$ observed in frame t.

To get a sense for the magnitude of p, we use data from a similar experiment: in this case, each origami has been designed in such a way that it carries exactly one fluorophore. We estimate p as the average ratio $\hat{p}$ of the number of frames where the fluorophore is "on" (a bright spot is seen) and the total number of observed frames before bleaching (no spot is seen for any time in the future), and we find $\hat{p} \approx 0.0339$. This indicates that we are in the difficult "small p" regime of the $\mathrm{Bin}(n, p)$ problem, and we will therefore apply the Bayesian estimators introduced in Section 3 (SE, DGE) to estimate $n^{(t)}$. The beta prior for SE and DGE uses the parameters $a = 2$ and $b = 2/\hat{p} - 2 \approx 56.99$. We choose the unimodal prior with $a = 2$, as suggested by Table 2, since we assume that our guess $\hat{p}$ is reasonably accurate. Note that a finer degree of modeling would require to view $n^{(0)}$, $n^{(t)}$ and p as random variables (with small variances) instead of constants. However, as shown in Section 3, the Bayesian estimators we consider are robust against fluctuations in the parameters and are therefore suited to estimate the respective mean values.

Since most fluorophores are deliberately forced to be "on" in the first frame, the relation $X^{(t)} \sim \mathrm{Bin}(n^{(t)}, p)$ does not hold initially. It only becomes valid after the initial state has relaxed to an equilibrium, which is why we only take into account data after frame 1500 (≈ 22.5 seconds). To mitigate the influence of correlations between observations (i.e., $X^{(t)}$ and $X^{(t+1)}$ for a spot cannot be considered independent), we also add a waiting time of 1500 frames between the frames we use for our analysis. In total, we use the six frames at six time points with $t \in \{1500, 3000, 4500, 6000, 7500, 9000\}$ depicted in Figure 2. The 94 realizations of $X^{(t)}$ are extracted from the image data as follows: at each registered origami position, represented by a 6×6 pixel ROI, the total brightness is measured and then divided by the brightness of a single fluorophore. We determined the brightness of a single fluorophore from the late frames of the experiment, where at most one fluorophore of each origami is active with high probability.

The results for the scale estimator with $\gamma = 0.5$ are depicted in Figure 3, which also shows the log-linear fit for model (4.1). This provides us with estimates for $n^{(0)}$ and B. The point estimates of $n^{(0)}$ and B for different estimators are summarized in Table 5. Given that it is to be expected that the true $n^{(0)}$ in this experiment lies between 13 and 16, we can see in Table 5 that the SEs with an improper prior ($\gamma \leq 1$) produce reasonable results, and the DGE also performs well. This confirms that we are indeed in the critical case of $\mathrm{Bin}(n, p)$ with small p, so that the prior putting a lot of weight on large values on n gives best results by correcting for the usual tendency to underestimate, see also the results of the simulation study performed under comparable conditions in Table 2. To illustrate the difficulty of this problem, Figure 4 shows exemplary counting results we obtained for $t \in \{1500, 7500\}$. Note

estimator	$n^{(0)}$	$B \cdot 10^3$
SE(0)	**16**	**0.152**
SE(0.5)	**13**	**0.148**
SE(1)	11	0.139
SE(2)	9	0.163
SE(3)	6	0.123
SE(5)	5	0.114
DGE(500)	**16**	**0.167**

TABLE 5: Estimates of the bleaching probability B and the number $n^{(0)}$ of fluorophore molecules on single DNA origami.

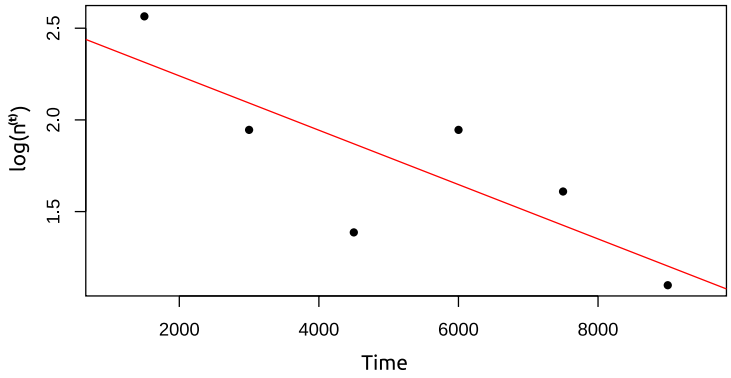


FIGURE 3: The log-linear fit described by $n^{(t)} = n^{(0)}(1-B)^t$ for the SE with $\gamma = 0.5$.

that the final estimates for $n^{(0)}$ are exclusively based on observations $X^{(t)} \leq 3$, where a great majority of these observations is already 0.

5 Proofs

5.1 Proof of the main theorems

Here we present the proofs of our posterior contraction results for n (Theorem 1 and 2). These require further technical results, e.g., fine moment estimates of a binomial random variable and bounds on the maximum of a triangular array of independent binomials, see Lemma 1 - 5. Auxiliary technicalities are postponed to the appendix.

Throughout the proof of Theorem 1 we will be concerned with an exemplary sequence in $\mathcal{M}_\lambda$. We call this sequence $(n_k, p_k)_k$ instead of $(n_k^0, p_k^0)_k$ for notational simplification. Our

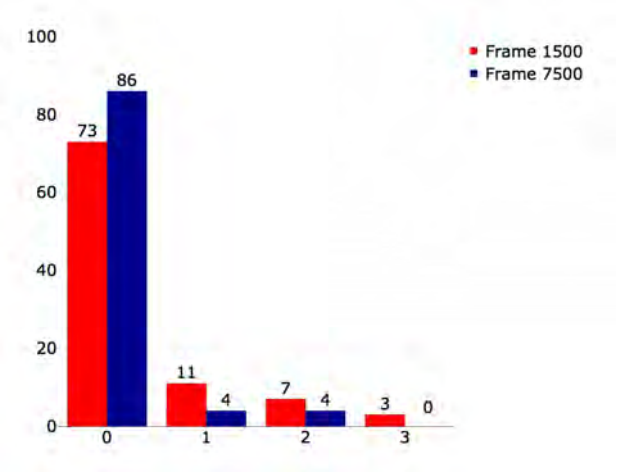


FIGURE 4: Bar charts of the observed numbers of fluorophore molecules for time frames 1500 and 7500.

arguments do not depend on the specific choice of $(n_k, p_k)_k$ but only rely on the parameters λ, a and b.

Proof of Theorem 1. First observe that

$$\Pi\big(N \neq n_k \,|\, X^k\big) = \frac{\sum_{n\neq n_k, n\geq M_k} L_{a,b}(n)\,\Pi_N(n)}{\sum_{n=M_k}^{\infty} L_{a,b}(n)\,\Pi_N(n)} \leq \sum_{n\neq n_k, n\geq M_k} \frac{L_{a,b}(n)\,\Pi_N(n)}{L_{a,b}(n_k)\,\Pi_N(n_k)}.$$

Under the assumption that $S_k \geq 2$ (which we justify below), we can apply Lemma 6 and find

$$\frac{L_{a,b}(n)}{L_{a,b}(n_k)} \leq c_1 k n_k \, \frac{L_{\lfloor a\rfloor,b}(n)}{L_{\lfloor a\rfloor,b}(n_k)}$$

for $c_1 = 1 + \lceil a\rceil + b$, where $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ denote the ceil and floor functions, respectively. It follows that

$$\Pi\big(N \neq n_k \,|X^k\big) \leq c_1 k n_k \sum_{n\neq n_k, n\geq M_k} \exp\left(k \int_{n_k}^{n} f'(m)\,\mathrm{d}m\right) \frac{\Pi_N(n)}{\Pi_N(n_k)} \tag{5.1}$$

with $f(m) = \frac{1}{k}\log L_{\lfloor a\rfloor,b}(m)$. If $n < n_k$, we can write $\int_{n_k}^{n} f'(m)\,\mathrm{d}m = -\int_{n}^{n_k} f'(m)\,\mathrm{d}m$. For an upper bound on the posterior we thus need a lower bound of $f'(m)$ if $m \leq n_k$ and an upper bound if $m \geq n_k$. Since f only depends on a via $\lfloor a\rfloor$, we assume that $a \in \mathbb{N}_0$ in the following. Then we can apply Lemma 4.1 from Hall (1994) and find

$$f'(m) = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{X_i} \frac{1}{m-j+1} - \sum_{j=1}^{S_k+a} \frac{1}{km+a+b-j} = \sum_{j=1}^{M_k} \frac{T_j - U_j}{j} - \sum_{j=M_k+1}^{S_k+a} \frac{U_j}{j} \tag{5.2}$$

with

$$T_j := \frac{1}{k}\sum_{i=1}^{k} \frac{(X_i)_j}{(m)_j} \qquad \text{and} \qquad U_j := \frac{(S_k + a)_j}{(km + a + b - 1)_j}$$

for $j \leq M_k$ and $j \leq S_k + a$ respectively, where $(t)_j = t\,(t-1)\cdots(t-j+1)$ denotes the falling factorial for $t > 0$. For convenience we define $T_j := 0$ for all $j > M_k$. Next, we introduce the events

$$\begin{aligned}
\mathcal{U}_k &:= \Big\{ M_k = n_k \;\text{ or }\; M_k \geq l_k \Big\}, \qquad \mathcal{R}_k := \Big\{ M_k \leq 2\log(k) \Big\}, \\
\mathcal{T}_{kj} &:= \Big\{ (m)_j \big| T_j - \mathbb{E}T_j \big| \leq \sqrt{(c_2 j)^j \, l_k \log(k)/k} \Big\}, \qquad \mathcal{T}_k := \bigcap_{j \in \mathbb{N}} \mathcal{T}_{kj}, \\
\mathcal{S}_k &:= \Big\{ \big| S_k - k n_k p_k \big| \leq \sqrt{\lambda\, k \log(k)} \Big\},
\end{aligned}$$

and denote the intersection $\mathcal{U}_k \cap \mathcal{R}_k \cap \mathcal{T}_k \cap \mathcal{S}_k$ by $\mathcal{A}_k$. The constant $c_2 = 2\,\lambda\,(\lambda + 2)$ is chosen to satisfy Lemma 2 for each sequence $(n_k, p_k)_k \in \mathcal{M}_\lambda$, and l_k is a fixed sequence with $l_k = o(\sqrt{\log(k)})$. Note that the sets $\mathcal{T}_{kj}$ are in fact independent of m due to the definition of T_j. On the event $\mathcal{S}_k$, Lemma 8 grants us the additional property

$$\big| U_j - \widetilde{U}_j \big| \leq j \sqrt{\frac{\lambda \log(k)}{k}} \left(\frac{c_3}{m} \right)^j \qquad \text{with} \qquad \widetilde{U}_j := \frac{(k n_k p_k + a)_j}{(km + a + b - 1)_j}$$

for $j \leq S_k + a$ and $c_3 = 2e^2(3\lambda + a + 1)$. Also note that $S_k \leq 2k\lambda$ holds and that $S_k \geq 2$ is guaranteed for $k/\lambda - \sqrt{\lambda\, k \log(k)} \geq 2$ on $\mathcal{S}_k$. Thus, equations (5.1) and (5.2) apply on $\mathcal{A}_k$ if k is sufficiently large.

Indeed, we can restrict our attention to $\mathcal{A}_k$, since

$$\mathbb{E}_{n_k,p_k}\big[\Pi\big(N \neq n_k \,|\, X^k \big) \big] - \mathbb{E}_{n_k,p_k}\big[\mathbf{1}_{\mathcal{A}_k} \Pi\big(N \neq n_k \,|\, X^k \big) \big] \leq \mathbb{P}_{n_k,p_k}\big(\mathcal{A}_k^{\mathrm{c}} \big) \longrightarrow 0 \tag{5.3}$$

for $k \to \infty$. To see this, one can bound $\mathbb{P}_k\big(\mathcal{A}_k^{\mathrm{c}}\big) := \mathbb{P}_{n_k,p_k}\big(\mathcal{A}_k^{\mathrm{c}}\big)$ by

$$\mathbb{P}_k\big(\mathcal{A}_k^{\mathrm{c}}\big) \leq \mathbb{P}_k\big(\mathcal{S}_k^{\mathrm{c}}\big) + \mathbb{P}_k\big(\mathcal{U}_k^{\mathrm{c}}\big) + 2\,\mathbb{P}_k\big(\mathcal{R}_k^{\mathrm{c}}\big) + \mathbb{P}_k\big(\mathcal{T}_k^{\mathrm{c}} \cap \mathcal{R}_k\big).$$

The first contribution vanishes due to Chebyshev's inequality (see e.g., DeGroot and Schervish (2012)), and the second and third terms are controlled by Lemma 3 and Lemma 4 respectively. The last contribution satisfies

$$\mathbb{P}_k\big(\mathcal{T}_k^{\mathrm{c}} \cap \mathcal{R}_k\big) = \mathbb{P}_k\left(\bigcup\nolimits_{j=1}^{\lfloor 2\log(k) \rfloor} \mathcal{T}_{kj}^{\mathrm{c}} \right) \leq \frac{\lfloor 2\log(k) \rfloor}{l_k \log(k)} \longrightarrow 0$$

for $k \to \infty$ due to Lemma 2. It is important to note that the upper bounds in these considerations only depend on λ if suitable choices for the involved constants are made. In the following, we always assume that $X^k \in \mathcal{A}_k$.

Auxiliary lower bound. For $M_k \le m < n_k$, we prove a lower bound of $f'(m)$. We may assume that $M_k \ge l_k \to \infty$ for $k \to \infty$ in this case, since $X^k \in \mathcal{U}_k$. For k such that $l_k \ge 4$ we can bound equation (5.2) by

$$f'(m) \ge \sum_{j=1}^{4} \frac{T_j - U_j}{j} - \sum_{5}^{S_k+a} \frac{U_j}{j}, \tag{5.4}$$

as $T_j \ge 0$ for all j. In case of $j = 1$ we obtain

$$T_1 - U_1 = \frac{S_k}{km} - \frac{S_k + a}{km + a + b - 1} \ge -\frac{a+1}{km-1} \ge -2\,\frac{(a+1)}{km} \ge -2\,\frac{\lambda(a+1)}{m^2}\sqrt{\frac{\log(k)}{k}},$$

where we used the upper bound $m < n_k \le \lambda\sqrt{k\log(k)}$ in the last inequality, which is guaranteed in $\mathcal{M}_\lambda$. To handle the terms with $j \ge 2$, we exploit that $X^k \in \mathcal{T}_k$ and apply $(m)_j \ge \big(m/e^2\big)^j$ from Lemma 7 in order to derive

$$\sum_{j=2}^{4} \frac{\big|T_j - \mathbb{E}T_j\big|}{j} \le \sqrt{\frac{l_k \log(k)}{k}} \sum_{j=2}^{4} \left(\frac{\sqrt{c_2 j}}{m/e^2}\right)^j \le 2\,\frac{4\,c_2\,e^4}{m^2}\,\sqrt{\frac{l_k\log(k)}{k}},$$

for k large enough such that $\sqrt{4\,c_2}\,e^2/l_k < 1/2$. Similarly, we find

$$\sum_{j=2}^{S_k+a} \frac{\big|U_j - \widetilde{U}_j\big|}{j} \le \sqrt{\frac{\lambda\log(k)}{k}} \sum_{j=2}^{S_k+a} \left(\frac{c_3}{m}\right)^j \le 2\,\frac{\sqrt{\lambda}\,c_3^2}{m^2}\,\sqrt{\frac{\log(k)}{k}}$$

and

$$\sum_{j=5}^{S_k+a} \frac{\widetilde{U}_j}{j} \le \sum_{j=5}^{S_k+a} \frac{1}{j}\left(\frac{c_4}{m}\right)^j \le 2\,\left(\frac{c_4}{m}\right)^5$$

for k (and thus m) sufficiently large. In the latter equation, we applied the first part of Lemma 9 with $c_4 = 6e^2(\lambda + a)$, using $S_k \le 2\,k\lambda$ on $\mathcal{S}_k$. Next, we note that $\mathbb{E}(X_1)_j = (n_k)_j\,p^j$ and consult the first result of Lemma 10 to establish

$$\sum_{j=2}^{4} \frac{\mathbb{E}T_j - \widetilde{U}_j}{j} \ge \frac{1}{2\lambda^2}\frac{n_k - m}{n_k m^3} - \frac{3\,c_5^4}{k} \ge \frac{1}{2\lambda^2}\frac{n_k - m}{n_k m^3} - 2\,\frac{2\,\lambda^2 c_5^4}{m^2}\sqrt{\frac{\log(k)}{k}},$$

where $c_5 = 3\,\lambda + 2\,a + 2$. Similar to the case $j = 1$, we applied $m^2 < \lambda^2\sqrt{k\log(k)}$ in the last inequality. All bounds calculated above can be inserted into inequality (5.4), yielding

$$\begin{aligned} f'(m) &\ge (T_1 - U_1) + \sum_{j=2}^{4} \frac{T_j - \mathbb{E}T_j}{j} + \sum_{j=2}^{S_k+a} \frac{\widetilde{U}_j - U_j}{j} + \sum_{j=2}^{4} \frac{\mathbb{E}T_j - \widetilde{U}_j}{j} - \sum_{j=5}^{S_k+a} \frac{\widetilde{U}_j}{j} \\ &\ge \frac{1}{4\lambda^2}\frac{n_k - m}{n_k m^3} - \frac{C_2}{m^2}\sqrt{\frac{l_k\log(k)}{k}} + \left[\frac{1}{4\lambda^2}\frac{n_k - m}{n_k m^3} - 2\,\left(\frac{c_4}{m}\right)^5\right] \\ &\ge \frac{C_1}{m^3}\frac{n_k - m}{n_k} - \frac{C_2}{m^2}\sqrt{\frac{l_k\log(k)}{k}} + h(m), \end{aligned} \tag{5.5}$$

where $h(m)$ is defined as the expression in square brackets. The constants in this bound are $C_1 = 1/4\lambda^2$ and $C_2 = 2\,\big(\lambda(a+1) + 4\,c_2\,e^4 + \sqrt{\lambda}\,c_3^2 + 2\,\lambda^2 c_5^4\big)$.

Auxiliary upper bound. We next provide an upper bound for $f'(m)$ if $m > n_k \geq M_k$. Unlike for the lower bound, we cannot assume that m becomes larger than any given constant with increasing k. Since U_j is nonnegative, we can bound

$$f'(m) \leq \sum_{j=1}^{M_k} \frac{T_j - U_j}{j}.$$

in equation (5.2). We look at the case $j = 1$ first, and see

$$T_1 - U_1 = \frac{S_k}{km} - \frac{S_k + a}{km + a + b - 1} \leq \frac{S_k\,(a+b)}{km\,(km-1)} \leq \frac{4\lambda\,(a+b)}{km^2} \leq \frac{4\lambda\,(a+b)}{m^2}\sqrt{\frac{\log^3(k)}{k}},$$

where we used that $S_k \leq 2\lambda\,k$ on the event $\mathcal{S}_k$. Next we set $\widetilde{m} := 4\,c_2\,e^4$ and derive

$$\begin{aligned}
\sum_{j=2}^{M_k} \frac{\left|T_j - \mathbb{E}T_j\right|}{j} &\leq \sqrt{\frac{l_k \log(k)}{k}} \sum_{j=2}^{M_k} \left(\frac{\sqrt{c_2 j}}{m/e^2}\right)^j \\
&\leq \frac{c_2 M_k\,e^4}{m^2} \sqrt{\frac{l_k \log(k)}{k}} \sum_{j=0}^{\lfloor m \rfloor} \left(\frac{e^2\sqrt{c_2}}{\sqrt{m}}\right)^j \\
&\leq \frac{c_2 M_k\,e^4}{m^2} \sqrt{\frac{l_k \log(k)}{k}} \cdot \begin{cases} 2 & \text{if } m > \widetilde{m} \\ \widetilde{m}\,(e^2\sqrt{c_2}+1)^{\widetilde{m}} & \text{if } m \leq \widetilde{m} \end{cases} \\
&\leq \frac{c_6}{m^2} \sqrt{\frac{l_k \log^3(k)}{k}}
\end{aligned}$$

for $c_6 = 2\,c_2\,e^4\big(\widetilde{m}\,(e^2\sqrt{c_2}+1)^{\widetilde{m}} + 2\big)$. In the last step, we used that $M_k \leq 2\log(k)$ on the event $\mathcal{R}_k$. In a similar fashion, we can establish the bound

$$\sum_{j=2}^{M_k} \frac{\left|U_j - \widetilde{U}_j\right|}{j} \leq \sqrt{\frac{\lambda \log(k)}{k}} \sum_{j=2}^{M_k} \left(\frac{c_3}{m}\right)^j \leq \frac{c_7}{m^2} \sqrt{\frac{\log^3(k)}{k}},$$

where $c_7 = 4\sqrt{\lambda}\,c_3^2\,\big(c_3^{2c_3+1}+1\big)$. Finally, we apply the second claim of Lemma 10 and obtain

$$\sum_{j=2}^{M_k} \frac{\mathbb{E}T_j - \widetilde{U}_j}{j} \leq -C_1' \frac{m - n_k}{n_k m^3}$$

with $C_1' = \big(2\lambda^2\,(a+b+1)^2\,\big)^{-1}$ for sufficiently large k. We conclude

$$\begin{aligned}
f'(m) &\leq (T_1 - U_1) + \sum_{j=2}^{M_k} \frac{T_j - \mathbb{E}T_j}{j} + \sum_{j=2}^{M_k} \frac{\widetilde{U}_j - U_j}{j} + \sum_{j=2}^{M_k} \frac{\mathbb{E}T_j - \widetilde{U}_j}{j} \\
&\leq -\frac{C_1'}{m^3}\frac{m - n_k}{n_k} + \frac{C_2'}{m^2}\sqrt{\frac{l_k\,\log^3(k)}{k}} \qquad (5.6)
\end{aligned}$$

for $C_2' = 4\lambda\,(a+b) + c_6 + c_7$.

Posterior bound. By applying the two inequalities (5.5) and (5.6) for $m < n_k$ and $m > n_k$, we can now bound the posterior probability $\Pi\big(N \neq n_k \,|\, X^k\big)$ on the event $\mathcal{A}_k$ through equation (5.1). First, we observe that for $n \in \mathbb{N}$ with $n \neq n_k$

$$\int_{n_k}^{n} \frac{m-n_k}{n_k m^3}\,\mathrm{d}m = \frac{1}{2}\frac{(n-n_k)^2}{(n_k n)^2} \qquad \text{and} \qquad \int_{n_k}^{n} \frac{1}{m^2}\,\mathrm{d}m = \frac{1}{2}\frac{(n-n_k)^2}{(n_k n)^2}\frac{2n_k n}{n-n_k}.$$

It also holds for $n \neq n_k$ that

$$\left|\frac{n}{n_k} - 1\right| \geq \frac{1}{2n}\,. \tag{5.7}$$

Therefore, if $l_k \leq n < n_k$, the function $h(m)$ introduced in equation (5.5) satisfies

$$\begin{aligned}\int_{n}^{n_k} h(m)\,\mathrm{d}m &= \frac{C_1}{2}\,\frac{(n-n_k)^2}{(n_k n)^2} - \frac{c_4^5}{2}\,\frac{n_k^4 - n^4}{(n_k n)^4} \\ &\geq \frac{C_1}{2}\,\frac{(n-n_k)^2}{(n_k n)^2}\left(1 - \frac{4c_4^5}{C_1}\frac{1}{1-n/n_k}\frac{1}{n^2}\right) \geq 0\end{aligned} \tag{5.8}$$

for k such that $l_k \geq 8\,c_4^5/C_1$. Employing bound (5.5) thus yields

$$-k\int_{n}^{n_k} f'(m)\,\mathrm{d}m \leq -k\frac{C_1}{2}\frac{(n-n_k)^2}{(n_k n)^2}\left(1 - C\frac{n_k n}{n_k - n}\sqrt{\frac{l_k \log(k)}{k}}\right),$$

where the constant C is given by $2\,C_2/C_1$. On the other hand, for $n_k < n$, bound (5.6) similarly leads to

$$k\int_{n_k}^{n} f'(m)\,\mathrm{d}m \leq -k\frac{C_1'}{2}\frac{(n-n_k)^2}{(n_k n)^2}\left(1 - C'\frac{n_k n}{n-n_k}\sqrt{\frac{l_k \log^3(k)}{k}}\right)$$

for $C' = 2\,C_2'/C_1'$.

Finally, let $\widetilde{C}_1 = \min\big\{C_1, C_1'\big\}$ and $\widetilde{C} = \max\big\{C, C'\big\}$. We can apply inequality (5.7) (with n and n_k switched) to find for any $n \in \mathbb{N}$ with $n \neq n_k$ and $n \geq M_k$ that

$$\begin{aligned}k\int_{n_k}^{n} f'(m)\,\mathrm{d}m &\leq -k\frac{\widetilde{C}_1}{2\,n_k^2}\left(\frac{n_k}{n} - 1\right)^2\left(1 - \frac{\widetilde{C}\,n_k}{|1-n_k/n|}\sqrt{\frac{l_k \log^3(k)}{k}}\right) \\ &\leq -k\frac{\widetilde{C}_1}{8\,n_k^4}\left(1 - 2\,\widetilde{C}\,n_k^2\sqrt{\frac{l_k \log^3(k)}{k}}\right) \leq -\frac{\widetilde{C}_1}{16}\frac{k}{n_k^4}\end{aligned}$$

for k large enough such that $n_k^2 \leq \sqrt{k/\big(l_k \log^3(k)\big)}/4\widetilde{C}$ for each sequence in $\mathcal{M}_\lambda$. Consulting inequality (5.1) and using the constraint $\Pi_N(n_k) \geq \beta \exp\big(-\alpha n_k^2\big)$ of the prior yields

$$\mathbf{1}_{\mathcal{A}_k}\Pi\big(N \neq n_k \,|X^k\big) \leq c_1 k n_k \sum_{n \neq n_k} \exp\left(-\frac{\widetilde{C}_1}{16}\frac{k}{n_k^4}\right)\frac{\Pi_N(n)}{\Pi_N(n_k)} \tag{5.9}$$

$$\leq \frac{c_1}{\beta} \exp\left(-\frac{\widetilde{C}_1}{16} \frac{k}{n_k^4} + \alpha n_k^2 + \log\left(k n_k\right) \right)$$
$$\leq \frac{c_1}{\beta} \exp\left(-\frac{\widetilde{C}_1}{16\,\lambda^4}\, k^{1/3} \log(k)^{2/3} + \alpha\lambda^2 \frac{k^{1/3}}{\log(k)^{1/3}} + \log\left(\lambda k^2\right) \right) \longrightarrow 0$$

for $k \to \infty$. Due to statement (5.3) this is sufficient to prove Theorem 1. □

Proof of Theorem 2. The result follows from the proof of Theorem 1, where we only need to handle the inequalities in (5.9) differently. Bounding the sum over the prior in (5.9) by an integral, one sees that the upper bound on T_k in (2.1) is sufficient to ensure convergence towards 0, if $\Pi_{N,k}$ is considered instead of Π_N. □

5.2 Auxiliary results for binomial random variables

Lemma 1. *Let X be a binomial random variable, $X \sim \mathrm{Bin}(n, p)$, for $n \in \mathbb{N}$ and $p \in (0, 1)$. Then*

$$\mathbb{E}[X^r] \leq B_r \cdot \max\{np, (np)^r\},$$

where B_r is the r-th Bell number.

Proof. Let $q = (1 - p)$ and let $M_{n,p}$ be the moment generating function of the binomial distribution,

$$M_{n,p}(t) = (pe^t + q)^n = f(g(t)),$$

where $f(s) = s^n$ and $g(t) = pe^t + q$. To obtain the moments of X, we look at the derivatives of $M_{n,p}$ at $t = 0$. The r-th derivatives of f and g are

$$f^{(r)}(s) = (n)_r\, s^{n-r} \quad \text{and} \quad g^{(r)}(t) = pe^t$$

for $r \in \mathbb{N}$. Since $g(0) = 1$, it holds that $f^{(r)}\big(g(0)\big) = (n)_r$. Furthermore, $g^{(r)}(0) = p$ for all r. We employ the Bell polynomial version of Faà di Bruno's formula, see Johnson (2002) equation (2.2), which is

$$(f \circ g)^{(r)}(t) = \sum_{k=1}^{r} f^{(k)}\big(g(t)\big)\, B_{r,k}\left(g^{(1)}(t), g^{(2)}(t), \ldots, g^{(r-k+1)}(t)\right). \tag{5.10}$$

The Bell polynomials $B_{r,k}$ are homogeneous of degree k. Therefore,

$$\mathbb{E}[X^r] = (f \circ g)^{(r)}(0) = \sum_{k=1}^{r} f^{(k)}(g(0))\, B_{r,k}\left(g^{(1)}(0), \ldots, g^{(r-k+1)}(0)\right)$$

$$
\begin{aligned}
&= \sum_{k=1}^{r} B_{r,k}(1,\dots,1)\,(n)_k p^k \\
&\leq B_r \cdot \max\{np, (np)^r\},
\end{aligned}
$$

where $B_r = \sum_{k=1}^{r} B_{r,k}$ is the r-th Bell number. □

Lemma 2. *Let $n \in \mathbb{N}$ and $p \in (0,1)$. Define k i.i.d. binomial random variables $X_1, \dots, X_k$ with distribution $\mathrm{Bin}(n,p)$. For each $j \in \mathbb{N}$ with $j \leq n$, the inequality*

$$
\mathbb{P}\left(\left| \frac{1}{k} \sum_{i=1}^{k} (X_i)_j - \mathbb{E}[(X_1)_j] \right| > \sqrt{\frac{l\,(cj)^j}{k}} \right) \leq \frac{1}{l}
$$

holds for any $l > 0$ and $c \geq 2\,np\,(np+2)$.

Proof. We define the random variable $\widetilde{X} \sim \mathrm{Bin}(n-j,p)$ and note that $\mathbb{E}[(X_i)_j] = (n)_j p^k$. Invoking Lemma 1, we derive the upper bound

$$
\begin{aligned}
\mathrm{Var}[(X_i)_j] &\leq \mathbb{E}[(X_i)_j^2] \\
&\leq (n)_j p^j\, \mathbb{E}[(\widetilde{X}+j)_j] \\
&\leq (np)^j \mathbb{E}[(\widetilde{X}+j)^j] \\
&\leq 2^j (np)^j \left(\mathbb{E}[\widetilde{X}^j] + j^j \right) \\
&\leq 2^j (np)^j \left(B_j\,(np+1)^j + j^j \right) \\
&\leq \left(2j\,np\,(np+2) \right)^j
\end{aligned}
$$

on the variance of $(X_i)_j$. The second inequality becomes transparent from expanding the expectation as a sum, and the last inequality is valid due to the relation $B_j \leq j^j$ that can be found in Berend and Tassa (2010). For $c \geq 2\,np\,(np+2)$, we obtain by Chebyshev's inequality that

$$
\mathbb{P}\left(\left| \frac{1}{k} \sum_{i=1}^{k} (X_i)_j - \mathbb{E}[(X_1)_j] \right| > \sqrt{\frac{l\,(cj)^j}{k}} \right) \leq \frac{\mathrm{Var}[(X_1)_j]/k}{l\,(cj)^j/k} \leq \frac{1}{l}.
$$

□

Lemma 3. *Let $n \in \mathbb{N}$ and $p \in (0,1)$, and let M_k denote the maximum of k independent binomial variables $X_1, \dots, X_k \sim \mathrm{Bin}(n,p)$. Let $(l_k)_{k\in\mathbb{N}}$ be such that $l_k \to \infty$ and $l_k = o(\sqrt{\log(k)})$. Then, for each k with $l_k > \max\{1, 4\,np\}$,*

$$
\mathbb{P}\big(M_k < \min\{l_k, n\}\big) \leq e^{-d_k},
$$

where

$$
d_k = \min\left\{ \frac{k}{e^{l_k \log(l_k/np)}}, \frac{k\,np}{8\pi\, l_k^2 e^{l_k^2/np}} \right\} \to +\infty \quad \text{as } k \to \infty.
$$

Proof. We have that

$$\mathbb{P}\big(M_k < \min\{l_k, n\}\big) = \begin{cases} \mathbb{P}(M_k < n) & \text{if } n \le l_k, \\ \mathbb{P}(M_k < l_k) & \text{if } n > l_k. \end{cases}$$

In case of $l_k \ge n$, we derive the upper bound

$$\log \mathbb{P}(M_k < n) \le -kp^n \le -k\, e^{-l_k \log(l_k/np)} \longrightarrow -\infty, \text{ as } k \to \infty,$$

by applying Bernoulli's inequality and for $l_k = o(\sqrt{\log(k)})$. If $n > l_k$, we find that $p \le 1/4$, and thus Slud's bound from Telgarsky (2010) can be applied to yield

$$\begin{aligned} \mathbb{P}(M_k < l_k) &= \big(1 - \mathbb{P}(X_1 \ge l_k)\big)^k \\ &\le \Phi\left(\frac{l_k}{\sqrt{np(1-p)}}\right)^k \le \Phi\left(\frac{\sqrt{2}\, l_k}{\sqrt{np}}\right)^k, \end{aligned}$$

where Φ is the cumulative standard normal distribution function. By the lower tail bound in Gordon (1941), which states that $1 - \Phi(t) \ge \frac{1}{2\pi}\frac{t}{t^2+1}\, e^{-t^2/2}$ for $t > 0$, we obtain

$$\begin{aligned} \Phi\left(\frac{\sqrt{2}\, l_k}{\sqrt{np}}\right)^k &\le \left(1 - \frac{np}{8\pi\, l_k^2}\, e^{-\frac{l_k^2}{np}}\right)^k \\ &\le \exp\left(-\frac{k}{8\pi}\, \frac{np}{l_k^2\, e^{l_k^2/np}}\right) \longrightarrow 0, \text{ as } k \to \infty, \end{aligned}$$

where we set $t = \sqrt{2}\, l_k/\sqrt{np} > 1$ (by assumption) and used $t/(t^2+1) \ge 1/2t^2$ for $t \ge 1$ and $l_k = o\left(\sqrt{\log(k)}\right)$. □

Lemma 4. *Let $n \in \mathbb{N}$ and $p \in (0,1)$. Define k i.i.d. binomial random variables $X_1, \ldots, X_k$ with distribution* $\mathrm{Bin}(n,p)$*, and let $M_k := \max_{i=1,\ldots,k} X_i$. Then*

$$\mathbb{P}\big(M_k \le 2\log(k)\big) \ge \left(1 - \frac{1}{k^2}\right)^k$$

if $k \ge e^{3np}$. Consequently, $\mathbb{P}\big(M_k > 2\log(k)\big) \to 0$, as $k \to \infty$.

Proof. We can write X_1 as a sum of n i.i.d. Bernoulli random variables bounded by 1. By Bernstein's inequality (see e.g., van der Vaart and Wellner (1996))

$$\begin{aligned} \mathbb{P}\left(M_k \le 2\log(k)\right) &= \big(1 - \mathbb{P}\big(X_1 - np > 2\log(k) - np\big)\big)^k \\ &\ge \left(1 - \exp\left\{-\frac{\big(2\log(k) - np\big)^2}{2\big(np(1-p) + \log(k)/3\big)}\right\}\right)^k \\ &\ge \big(1 - e^{-2\log(k)}\big)^k, \end{aligned}$$

where the last inequality holds for $\log(k) \ge 3np$. □

Lemma 5. *Let $(n_k, p_k)_{k\in\mathbb{N}}$ be a sequences with $n_k \in \mathbb{N}$, $p_k \in (0,1)$ and $n_k p_k \to \mu > 0$. Define the independent random variables $X_1, \dots, X_k \sim \mathrm{Bin}(n_k, p_k)$ and let $M_k := \max_{i=1,\dots,k} X_i$.*

(i) *If $n_k \log(n_k) < c \log(k)$ for $c < 1$, then $\mathbb{P}(M_k = n_k) \to 1$, as $k \to \infty$.*
(ii) *If $n_k \log(n_k) > c \log(k)$ for $c > 1$, then $\mathbb{P}(M_k = n_k) \to 0$, as $k \to \infty$.*

Proof. (i): We have convergence of the sample maximum towards the parameter n_k if

$$\mathbb{P}(M_k = n_k) = 1 - \big(1 - p_k^{n_k}\big)^k \geq 1 - e^{-k p_k^{n_k}} \to 1, \text{ as } k \to \infty,$$

where we applied Bernoulli's inequality. This holds if $\log(k) - n_k \log(n_k/n_k p_k) \to \infty$, which follows from

$$\frac{n_k \log(n_k)}{\log(k)} < c < 1 \qquad \text{and} \qquad \frac{n_k \,|\log(n_k p_k)|}{\log(k)} \leq \frac{c\,|\log(n_k p_k)|}{\log\big(c \log(k)\big)}.$$

(ii): It holds that $P(M_k = n_k) \leq k p_k^{n_k} \leq \exp\big(\log(k) - n_k \log(n_k/n_k p_k)\big)$. Similar to the argument above, the right hand side in this inequality converges to 0 since

$$\frac{n_k \log(n_k)}{\log(k)} > c > 1 \qquad \text{and} \qquad \frac{n_k \,|\log(n_k p_k)|}{\log(k)} \leq \frac{c\,|\log(n_k p_k)|}{\log\big(c \log(k)\big)}.$$

□

A Auxiliary technicalities

Lemma 6. *For $k, n, s \in \mathbb{N}$ and $b > 0$ such that $2 \leq s \leq kn$ define the function*

$$f(a) = \frac{\Gamma\big(kn - s + b\big)\Gamma\big(s + a\big)}{\Gamma(kn + a + b)}$$

for $a \geq 0$. Then f is monotonically decreasing and $f\big(\lfloor a \rfloor\big)/f\big(\lceil a \rceil\big) \leq c\,kn$ for $c \geq 1 + \lceil a \rceil + b$.

Proof. It is sufficient to look at $h(a) := \Gamma\big(y + a\big)/\Gamma(z + a)$, where $2 \leq y < z$ are fixed. For $\epsilon > 0$, we find that $\log h(a + \epsilon) \leq \log h(a)$ is equivalent to

$$\gamma(y + a + \epsilon) - \gamma(y + a) \leq \gamma(z + a + \epsilon) - \gamma(z + a)$$

with $\gamma(t) = \log \Gamma(t)$ for $t > 0$. This inequality is true since γ is convex, see Merkle (1996), which therefore establishes monotonicity. We also find

$$\frac{h\big(\lfloor a \rfloor\big)}{h\big(\lceil a \rceil\big)} = \frac{z + \lceil a \rceil - 1}{y + \lfloor a \rfloor - 1} \leq z + \lceil a \rceil.$$

Substituting $y = s$ and $z = kn + b$, and using that $kn + \lceil a \rceil + b \leq (1 + \lceil a \rceil + b)\, kn$ yields the second result. □

Lemma 7. *Let $j \in \mathbb{N}$ and $n, m > 1$ with $j \leq \min\{m, n\}$. Let $(a)_j = a(a-1)\dots(a-j+1)$ denote the falling factorial for $a \in \mathbb{R}$.*

1. *For $0 < c \leq e^{-2}$ it holds that $(c\,m)^j \leq (m)_j \leq m^j$.*
2. *For $n \geq m$ and $j > 1$ it holds that $\frac{m^j (n)_j}{n^j (m)_j} \geq 1 + \frac{n-m}{nm}$.*

Proof. 1. From Theorem 1 in Jameson (2015) follows that

$$\sqrt{2\pi}\, m^{m+1/2} e^{-m} \leq \Gamma(m+1) \leq e\,\sqrt{2\pi}\, m^{m+1/2} e^{-m}\,.$$

We apply this to obtain

$$(m)_j = \frac{\Gamma(m+1)}{\Gamma(m-j+1)} \geq \frac{1}{e}\left(\frac{m}{m-j}\right)^{m-j+1/2} \left(\frac{m}{e}\right)^j \geq \left(\frac{m}{e^2}\right)^j.$$

2. For $n \geq m$ and $j > 1$, we bound

$$\frac{m^j (n)_j}{n^j (m)_j} = \prod_{i=0}^{j-1} \frac{m(n-i)}{(m-i)n} = \prod_{i=0}^{j-1} \left(1 + i\frac{n-m}{n(m-i)}\right) \geq 1 + \frac{n-m}{nm}.$$

□

Lemma 8. *Let $k, n, s \in \mathbb{N}$, $m > 0$ and $p \in (0,1)$ such that $k \geq 2$ and $km \geq s$. Let furthermore $a \geq 0, b > 0$ and define*

$$u_j = \frac{(s+a)_j}{(km+a+b-1)_j}, \qquad \tilde{u}_j = \frac{(knp+a)_j}{(km+a+b-1)_j}$$

for $j \in \mathbb{N}$ with $j \leq s + a$. Then it holds that

$$|s - knp| \leq \sqrt{\lambda\, k \log k} \quad \Longrightarrow \quad |u_j - \tilde{u}_j| \leq j \sqrt{\frac{\lambda \log k}{k}} \left(\frac{c}{m}\right)^j$$

for any $\lambda \geq np$ and $c \geq 2e^2\,(3\lambda + a + 1)$.

Proof. Let $t = \sqrt{\lambda\, k \log k}$ and assume that $|s - knp| =: |t'| \leq t$. Then, by applying a telescoping sum, we find

$$\begin{aligned} |u_j - \tilde{u}_j| &= \frac{|(knp + t' + a)_j - (knp + a)_j|}{(km+a+b-1)_j} \\ &\leq \sum_{l=0}^{j-1} \frac{|knp + t' + a| \dots |(knp + t' + a - l) - (knp + a - l)| \dots |knp + a - j + 1|}{(km+a+b-1)_j} \end{aligned}$$

$$\leq \sum_{l=0}^{j-1} \frac{(c_1 k)^{j-1}\, t}{(c_2 k m)^j} \leq j\, \frac{t}{k}\, \Big(\frac{c}{m}\Big)^j. \tag{A.1}$$

In the second inequality, we bound the numerator from above by noting that $t \leq k\lambda$ and thus $j \leq s + a \leq 2\,k\lambda + a$. Therefore,

$$knp + t + a + 1 + j \leq (3\lambda + a + 1)\, k =: c_1 k.$$

The denominator is bound from below by applying the first statement of Lemma 7, yielding

$$(km + a + b - 1)_j \geq \big((km-1)/e^2\big)^j \geq (c_2 k m)^j \tag{A.2}$$

for $c_2 = 1/2e^2$. In the final inequality of equation (A.1), c can be chosen as c_1/c_2 since $c_1 > 1$. □

Lemma 9. *Let $k, n \in \mathbb{N}$, $m > 0$ and $p \in (0,1)$ such that $k \geq 2$, and let $a \geq 0, b > 0$. For any $\lambda \geq np$ it holds that*

$$|\tilde{u}_j| := \left| \frac{(knp + a)_j}{(km + a + b - 1)_j} \right| \leq \Big(\frac{c_1}{m}\Big)^j$$

if $j \in \mathbb{N}$ with $j \leq 2\,k\lambda + a$ and $c_1 \geq 6e^2(\lambda + a)$. Furthermore, if $j \leq m$, then

$$\tilde{u}_j \leq \Big(\frac{np}{m}\Big)^j + \frac{j c_2^j}{k}$$

for any $c_2 \geq 3np + 2a + 2$.

Proof. The first result follows from bounding the numerator of $\tilde{u}_j$ from above by $\big(3k(\lambda+a)\big)^j$ and the denominator from below by equation (A.2) of the previous lemma. In case of $j \leq m$ it holds that

$$\begin{aligned} \frac{knp + a - i + 1}{km + a + b - i} - \frac{np}{m} &= \frac{m(a - i + 1) - np(a + b - i)}{m(km + a + b - i)} \\ &\leq \frac{inp + (a+1)m}{m(km - i)} \\ &\leq \frac{m(np + a + 1)}{m^2(k-1)} \leq 2\,\frac{np + a + 1}{k} =: \frac{\tilde{c}_2}{k} \end{aligned}$$

for each $i = 1, \ldots, j$. This inequality yields the upper bound

$$\tilde{u}_j \leq \left(\frac{np}{m} + \frac{\tilde{c}_2}{k}\right)^j.$$

We then apply the relation

$$(x + y)^j \leq x^j + j\, y\, \big(x + y\big)^{j-1}$$

for $x, y > 0$, which can be obtained from expanding $(x+y)^j$ as binomial sum, and conclude

$$\tilde{u}_j \le \left(\frac{np}{m}\right)^j + j\,\frac{\tilde{c}_2}{k}\left(\frac{np}{m} + \frac{\tilde{c}_2}{k}\right)^{j-1} \le \left(\frac{np}{m}\right)^j + \frac{jc_2^j}{k}$$

for $c_2 \ge \tilde{c}_2 + np$. □

Lemma 10. *Let $n, k \in \mathbb{N}$, $m > 0$ and $p \in (0,1)$ such that $k \ge 2$, and let $a \ge 0, b > 0$. Define*

$$t_j = \frac{(n)_j\, p^j}{(m)_j} \qquad \text{and} \qquad \tilde{u}_j = \frac{(knp+a)_j}{(km+a+b-1)_j}.$$

for $j \in \mathbb{N}$ with $1 < j \le m$. If $n > m$, it holds that

$$t_j - \tilde{u}_j \ge \frac{(np)^j}{n}\frac{n-m}{m^{j+1}} - \frac{jc_2^j}{k},$$

where $c_2 = 3np + 2a + 2$ from Lemma 9. If $n < m$, $j \le knp + a$, and $n \le \lambda \sqrt[4]{k}$ for some $\lambda > 0$, we also have

$$t_2 - \tilde{u}_2 \le -c\,\frac{m-n}{nm^3} \qquad \text{and} \qquad t_j - \tilde{u}_j \le 0$$

for $k \ge (1 + 1/np)^2\,(2\,\lambda\,(a+b+1))^4$ and $c \le \big(np/(a+b+1)\big)^2/2$.

Proof. Applying the respective second statements of Lemma 9 and Lemma 7, we establish

$$t_j - \tilde{u}_j \ge \left[\frac{(n)_j\, p^j}{(m)_j} - \left(\frac{np}{m}\right)^j\right] - \frac{jc_2^j}{k} \ge \left(\frac{np}{m}\right)^j \frac{n-m}{nm} - \frac{jc_2^j}{k}$$

for $n > m$, which shows the first result. For the second result, assume $m > n$. We look at the case $j = 2$ first. Direct calculation shows

$$\begin{aligned} t_2 - \tilde{u}_2 &= \frac{n(n-1)p^2}{m(m-1)} - \frac{(knp+a)(knp+a-1)}{(km+a+b-1)(km+a+b-2)} \\ &\le \frac{np}{m}\left(\frac{n-1}{m-1}p - \frac{np}{m}\frac{1-1/knp}{\big(1+(a+b)/km\big)^2}\right) \\ &\le -\frac{np}{nm}\frac{np\,(m-n) - \tilde{c}\,nm/k}{m\,(m-1)\,\big(1+(a+b)/km\big)^2} \end{aligned}$$

with $\tilde{c} = (1+np)\,(1+a+b)^2$. Under the assumed conditions, the numerator of the last expression can be bounded by

$$np\,(m-n) - \tilde{c}\,\frac{nm}{k} = np\,(m-n)\left(1 - \frac{\tilde{c}}{np}\frac{nm}{(m-n)k}\right) \ge \frac{np}{2}\,(m-n)$$

for $k \geq k_0 := (4\tilde{c}\lambda^2/np)^2$. This follows from

$$\frac{\tilde{c}}{np}\frac{nm}{(m-n)k} \leq \frac{\tilde{c}}{np}\frac{1}{k} \cdot \left\{ \begin{array}{ll} 2n & \text{if } m > 2n \\ 2n^2 & \text{if } m \leq 2n \end{array} \right\} \leq \frac{\tilde{c}}{np}\frac{2\lambda^2}{\sqrt{k}} \leq \frac{1}{2} \tag{A.3}$$

if $n \leq \lambda\sqrt[4]{k}$, and it implies that

$$t_2 - \tilde{u}_2 \leq -\frac{(np)^2}{2\,(a+b+1)^2}\frac{m-n}{nm^3} \leq 0.$$

Finally, for $2 \leq j \leq knp + a$ and $k \geq k_0$ we can derive

$$\frac{t_j}{\tilde{u}_j} = \frac{t_2}{\tilde{u}_2}\prod_{i=2}^{j-1}\frac{p\,(n-i)}{m-i}\,\frac{km+a+b-1-i}{knp+a-i} =: \frac{t_2}{\tilde{u}_2}\prod_{i=2}^{j-1}\frac{r_i}{v_i} \leq 1.$$

This statement is true due to $t_2/\tilde{u}_2 \leq 1$, and because $r_i \leq v_i$ is equivalent to

$$p(n-i)(a+b-1-i) - (m-i)(a-i) \leq ikp\,(m-n),$$

which follows from

$$np\,(a+b) + i\,(a+m) \leq i\,2\,\tilde{c}m \leq ikp\,(m-n).$$

The two inequalities hold because of the choice of $\tilde{c}$ and equation (A.3). □

Acknowledgements

Support of DFG CRC 755 (A6), Cluster of Excellence MBExC and DFG RTN 2088 (B4) is gratefully acknowledged. JSH was partially supported by a TOP II grant from the NWO. We are grateful to Oskar Laitenberger for providing us his data recorded at the Laser-Laboratorium Göttingen e.V.

Supplementary Material

Supplementary Movie: Fluorescence microscopy
The file provides a video of the first 9000 microscopic frames of the data set that was used for estimating the number of fluorophores in Section 4.
(doi: `http://www.stochastik.math.uni-Bgoettingen.de/SMS-Bmovie.mp4`)

References

Aspelmeier, T., A. Egner, and A. Munk (2015). Modern statistical challenges in high-resolution fluorescence microscopy. *Annual Review of Statistics and Its Application 2*(1), 163–202.

Basu, S. and N. Ebrahimi (2001). Bayesian capture-recapture models for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika 88*, 269–279.

Berend, D. and T. Tassa (2010). Improved bounds on bell numbers and on moments of sums of random variables. *Probability and mathematical statistics 30*, 185–205.

Berger, J. O., J. M. Bernardo, and D. Sun (2012). Objective priors for discrete parameter spaces. *Journal of the American Statistical Association 107*, 636–648.

Betzig, E., G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess (2006). Imaging intracellular fluorescent proteins at nanometer resolution. *Science 313*(5793), 1642–1645.

Blumenthal, S. and R. C. Dahiya (1981). Estimating the binomial parameter n. *Journal of the American Statistical Association 76*, 903–909.

Carroll, R. J. and F. Lombard (1985). A note on N estimators for the binomial distribution. *Journal of the American Statistical Association 80*, 423–426.

Castillo, I., J. Schmidt-Hieber, and A. W. van der Vaart (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics 43*, 1986–2018.

Castillo, I. and A. W. van der Vaart (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics 40*, 2069–2101.

DasGupta, A. and H. Rubin (2005). Estimation of binomial parameters when both n, p are unknown. *Journal of Statistical Planning and Inference 130*, 391–404.

DeGroot, M. H. and M. J. Schervish (2012). *Probability and Statistics*. Pearson Education.

Draper, N. and I. Guttman (1971). Bayesian estimation of the binomial parameter. *Technometrics 13*, 667–673.

Fisher, R. (1941). The negative binomial distribution. *Annals of Eugenics London 11*, 182–187.

Fölling, J., M. Bossi, H. Bock, R. Medda, C. A. Wurm, B. Hein, S. Jakobs, C. Eggeling, and S. W. Hell (2008). Fluorescence nanoscopy by ground-state depletion and single-molecule return. *Nature Methods 5*, 943–945.

Gao, C., A. W. van der Vaart, and H. H. Zhou (2015). A general framework for bayes structured linear models.

Ghosal, S., J. K. Ghosh, and A. W. van der Vaart (2000). Convergence rates of posterior distributions. *The Annals of Statistics 28*, 500–531.

Ghosal, S. and A. van der Vaart (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press.

Gordon, R. D. (1941). Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics 12*(3), 364–366.

Günel, E. and D. Chilko (1989). Estimation of parameter n of the binomial distribution. *Communications in Statistics - Simulation and Computation 18*, 537–551.

Haldane, J. B. S. (1941). The fitting of binomial distributions. *Annals of Human Genetics 11*, 179–181.

Hall, P. (1994). On the erratic behavior of estimators of N in the binomial N, p distribution. *Journal of the American Statistical Association 89*, 344–352.

Hamedani, G. G. and G. G. Walter (1988). Bayes estimation of the binomial parameter n. *Communications in Statistics - Theory and Methods 17*, 1829–1843.

Hell, S. W. (2009). Microscopy and its focal switch. *Nature Methods 6*, 24–32.

Hell, S. W. (2015). Nobel lecture: Nanoscopy with freely propagating light. *Reviews of Modern Physics 87*(4), 1169–1181.

Hess, S. T., T. P. Girirajan, and M. D. Mason (2006). Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophys J. 91*, 4258–72.

Jameson, G. (2015). A simple proof of stirling's formula for the gamma function. *The Mathematical Gazette 99*(544), 68.

Johnson, W. P. (2002). The curious history of Faà di Bruno's formula. *The Mathematical Association of America*.

Kahn, W. D. (1987). A cautionary note for bayesian estimation of the binomial parameter n. *The American Statistician 41*(1), 38–40.

Karathanasis, C., F. Fricke, G. Hummer, and M. Hellemann (2017). Molecule counts in localization microscopy with organic fluorophores. *ChemPhysChem 18*, 942–948.

Lee, S.-H., J. Y. Shin, A. Lee, and C. Bustamante (2012). Counting single photoavtiatable fluorescent molecules by photoactivated localization microscopy (palm). *PNAS 109 43*, 17436–17441.

Lehmann, E. and G. Casella (1996). *Theory of Point Estimation, 2 ed.* Springer.

Link, W. A. (2013). A cautionary note on the discrete uniform prior for the binomial n. *Ecology 94*, 2173–2179.

Merkle, M. (1996). Logarithmic convexity and inequalities for the gamma function. *Journal of mathematical analysis and applications 203*(2), 369–380.

Olkin, I., A. J. Petkau, and J. V. Zidek (1980). A comparison of n estimators for the binomial distribution. *Journal of the American Statistical Association 76*, 637–642.

Otis, D. L., K. P. Burnham, G. C. White, and D. R. Anderson (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs 64*, 1–135.

Raftery, A. E. (1988). Inference for the binomial N parameter: A hierachical Bayes approach. *Biometrika 75*, 223–228.

Rollins, G. C., J. Y. Shin, C. Bustamante, and S. Pressé (2015). Stochastic approach to the molecular counting problem in superresolution microscopy. *PNAS 112 2*, E110–8.

Royle, J. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics 60*, 108–115.

Rust, M. J., M. Bates, and X. Zhuang (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods 3*, 793–796.

Schmied, J. J., M. Raab, C. Forthmann, E. Pibiri, B. Wünsch, T. Dammeyer, and P. Tinnefeld (2014). Dna origami-based standards for quantitative fluoresence microscopy. *Nature Protocols 9*, 1367–1391.

Schneider, L. F., T. Staudt, and A. Munk (2018). Posterior consistency in the binomial (n, p) model with unknown n and p: A numerical study. preprint, arXiv:1809.02459.

Schwartz, L. (1965). On bayes procedures. *Z. Wahrsch. Verw. Gebiete 4*, 10–26.

Ta, H., J. Keller, M. Haltmeier, S. K. Saka, J. Schmied, F. Opazo, P. Tinnefeld, A. Munk, and S. W. Hell (2015). Mapping molecules in scanning far-field fluorescence nanoscopy. *Nature Communications 6*.

Telgarsky, M. (2010). Central binomial tail bounds. preprint, arXiv:0911.2077.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence*. Springer.

CHAPTER B

Posterior Consistency in the Binomial Model with Unknown Parameters: A Numerical Study

Posterior Consistency in the Binomial (n, p) Model with Unknown n and p: A Numerical Study

Laura Fee Schneider*[1], Thomas Staudt†[1], and Axel Munk‡[1,3]

[1]Institute for Mathematical Stochastics, University of Göttingen
[3]Max Planck Institute for Biophysical Chemistry, Göttingen

Abstract

Estimating the parameters from k independent $\mathrm{Bin}(n, p)$ random variables, when both parameters n and p are unknown, is relevant to a variety of applications. It is particularly difficult if n is large and p is small. Over the past decades, several articles have proposed Bayesian approaches to estimate n in this setting, but asymptotic results could only be established recently in [11]. There, posterior contraction for n is proven in the problematic parameter regime where $n \to \infty$ and $p \to 0$ at certain rates. In this article, we study numerically how far the theoretical upper bound on n can be relaxed in simulations without losing posterior consistency.

1 Introduction

We consider estimating the parameter n of the binomial distribution from k independent observations when the success probability p is unknown. This situation is relevant in many applications, for example in estimating the population size of a species [10] or the total number of defective appliances [4]. Another recent application is quantitative nanoscopy, see [11]. There, the total number of fluorescent markers (fluorophores) attached to so-called DNA-origami is estimated from a time series of microscopic images. The number of active fluorophores counted in each image is modeled as binomial observation, where the probability p that a fluorophore is active in the respective image is very small (often below 5%).

This setting, where the success probability p is small (and n potentially large), is very challenging. The difficulties that arise can be understood by considering the

*laura-fee.schneider@mathematik.uni-goettingen.de
†thomas.staudt@stud.uni-goettingen.de
‡munk@math.uni-goettingen.de

following property of the binomial distribution: if n converges to infinity, p converges to zero, and the product np converges to $\lambda > 0$, then a $\mathrm{Bin}(n,p)$ random variable converges in distribution to a Poisson variable with parameter λ. Thus, the binomial distribution converges to a distribution with a single parameter. This suggests that it gets harder to derive information about the two parameters separately when n is large and p small.

In this context, it is instructive to look at the sample maximum M_k as an estimator for n, which was suggested by Fisher in 1941 [5]. Although it turns out to be impractical, see [3], the sample maximum is consistent and converges in probability for fixed parameters (n,p) exponentially fast to the true n, as $k \to \infty$. This can be seen from

$$\mathbf{P}\left(M_k = n\right) = 1 - (1-p^n)^k, \tag{1}$$

which implies, by Bernoulli inequality and since $1 - x \leq e^{-x}$, that

$$1 - e^{-kp^n} \leq \mathbf{P}\left(M_k = n\right) \leq kp^n.$$

In an asymptotic setting where $n \to \infty$ and $p \to 0$ such that $kp^n \to 0$, the probability in (1) no longer converges to one. Thus, the sample maximum is a consistent estimator for n only as long as $kp^n \to \infty$. The condition $e^n = O(k)$ is necessary for this to hold.

Estimating n in this difficult regime becomes more manageable by including prior knowledge about p. We therefore consider random N and P, and variables $X_1, \ldots, X_k$ that are independently $\mathrm{Bin}(n,p)$ distributed given that $N = n$ and $P = p$. Various Bayesian estimators have been suggested over the last 50 years, see [4, 10, 1, 6, 7]. In all of this work, a product prior for (N,P) is used, and the prior Π_P on P is chosen as beta distribution $\mathrm{Beta}(a,b)$ for some $a,b > 0$. Since this is the conjugate prior, it is a natural choice. In contrast, there is quite some discussion about the most suitable prior Π_N for N, see for example [8, 9, 13, 1]. Therefore, the asymptotic results in [11] are described flexible in terms of Π_N, and they only require a condition that ensures that enough weight is put on large values of n (see equation (4) in Section 2).

In [11], we also introduce a new class of Bayesian point estimators for n, which we call scale estimators. We choose $\Pi_P \sim \mathrm{Beta}(a,b)$ and set $\Pi_N(m) \propto m^{-\gamma}$ for a positive value γ. If $\gamma > 1$, the prior Π_N is a proper probability distribution, but it is sufficient to ensure $\gamma + a > 1$ in order to obtain a well-defined posterior distribution. The scale estimator is then defined as the minimizer of the Bayes risk with respect to the relative quadratic loss, $l(x,y) = (x/y - 1)^2$. Following [10], it is given by

$$\hat{n} := \frac{\mathbb{E}\left[\frac{1}{N}|\mathbf{X}^k\right]}{\mathbb{E}\left[\frac{1}{N^2}|\mathbf{X}^k\right]} = \frac{\sum_{m=M_k}^{\infty} \frac{1}{m} L_{a,b}(m)\Pi_N(m)}{\sum_{m=M_k}^{\infty} \frac{1}{m^2} L_{a,b}(m)\Pi_N(m)}, \tag{2}$$

where $\mathbf{X}^k = (X_1, \ldots, X_k)$ denotes the sample, M_k is the sample maximum, and $L_{a,b}$ is the beta-binomial likelihood, see [2]. We refer to [11] for a detailed discussion and numerical study of this estimator.

The present article is structured as follows. In Section 2, the main theorem (proven in [11]) is presented, which shows uniform posterior contraction in the introduced Bayes setting for suitable asymptotics of n and p. The theorem states that $n^{6+\epsilon} = O(k)$ for $\epsilon > 0$ is already sufficient for consistency of the Bayes estimator, improving significantly over the sample maximum. In Section 3, we then conduct a simulation study to closer investigate the restrictions for the parameters n and p needed to ensure consistency. Our findings indicate that estimation of n is still consistent if $n^5 = O(k)$, but that it becomes inconsistent for $n^3 = O(k)$. It is hard to pin down the exact transition from consistency to inconsistency when $n^\alpha = O(k)$, but our results suggest that it happens close to $\alpha = 4$. We discuss our results and provide several remarks in Section 4.

2 Posterior Contraction for n

To study posterior contraction in the binomial model we consider the Bayesian setting described in Section 1. For fixed parameters n and p that are independent of the number of observations k, posterior consistency follows from Doob's theorem, see, e.g., [12]. We extend this result to the class of parameters

$$\mathcal{M}_\lambda := \left\{(n_k, p_k)_k : 1/\lambda \leq n_k p_k \leq \lambda,\ n_k \leq \lambda \sqrt[6]{k/\log(k)}\right\} \tag{3}$$

for fixed $\lambda > 1$. Since we want to handle a variety of suitable prior distributions for N, we only require that Π_N is a proper probability distribution on $\mathbb{N}$ that fulfills the condition

$$\Pi_N(m) \geq \beta e^{-\alpha m^2} \quad \forall\, m \in \mathbb{N} \tag{4}$$

for some positive constants α and β.

Theorem 1 (see [11]). *Conditionally on $N = n_k$ and $P = p_k$, let $X_1, \ldots, X_k \overset{i.i.d.}{\sim} \mathrm{Bin}(n_k, p_k)$. For any prior distribution $\Pi_{(N,P)} = \Pi_N \Pi_P$ on (N, P) with $\Pi_P = \mathrm{Beta}(a, b)$ for $a, b > 0$, and where Π_N satisfies (4), we have uniform posterior contraction over the set $\mathcal{M}_\lambda$ of sequences $(n_k, p_k)_k$ defined in (3) for any $\lambda > 1$, i.e.,*

$$\sup_{(n_k,p_k)_k \in \mathcal{M}_\lambda} \mathbb{E}_{n_k,p_k}\left[\Pi\big(N \neq n_k \mid \mathbf{X}^k\big)\right] \to 0, \text{ as } k \to \infty.$$

This result directly implies consistency of the scale estimator (2) for parameter sequences in $\mathcal{M}_\lambda$. The flexible restrictions on the prior distribution allow to apply the result to the estimators derived in [6] and [7] as well. Furthermore, it is possible to extend the statement of Theorem 1 to improper priors on N, as done in Theorem 2 in [11], in order to cover the estimators in [4] and [1].

3 Simulation Study

The theorem presented in the previous section states that the asymptotic behavior $n_k \sim O\big(\sqrt[6]{k/\log(k)}\big)$ leads to posterior contraction of N for suitable priors, as long as

$n_k p_k$ stays in a compact interval bounded away from zero. In this section we try to answer the question by how much the constraints on $\mathcal{M}_\lambda$ in Theorem 1 can be relaxed. We address this problem by studying the relation between posterior contraction and the order $\alpha > 0$ when $n_k \sim O\big(\sqrt[\alpha]{k}\big)$. More precisely, we are interested in the smallest $\alpha = \alpha^*$ such that the result

$$\mathbb{E}_{n_k,p_k}\left[\Pi\left(N \neq n_k \,|\, \mathbf{X}^k\right)\right] \to 0, \quad \text{as} \quad k \to \infty, \tag{5}$$

remains valid. Tackling this problem analytically turns out to be extremely challenging, see the proof of Theorem 1 in [11].

In our simulations, we consider sequences $(n_k, p_k)_k$ defined by $n_k = w\sqrt[\alpha]{k}$ and $p_k = \mu/n_k$ for parameters $w, \mu > 0$. The values of w and μ should, ideally, not matter for the asymptotics and thus for the pursuit of α^*. Suitable choices of w and μ for given α are still necessary for practical reasons to ensure that the asymptotic behavior becomes visible for the values of k covered by the simulations. For any selection (α, w, μ), we calculate the posterior probability of the true parameter n_k and the MSE of different estimators for values of k up to 10^{11}. In order to achieve these extremely large observation numbers, we take care to minimize the number of operations when expressing the beta-binomial likelihood $L_{a,b}$ in our implementation. Since $L_{a,b}$ does not depend on the order of the observations but only on the frequencies of each distinct outcome x_i, the runtime depends on n_k (the number of different values that x_i can take) instead of k itself.

Figures 1a–b show the (empirical) mean posterior probability in (5) and the (empirical) mean square error (MSE) between $\hat{n}$ and n for different scale estimators $\hat{n}$ in several scenarios (α, w, μ). The number of samples was set to 200. It is clearly visible that the choice $\alpha = 6$ leads to posterior consistency (which is in good agreement with Theorem 1), since the posterior probability approaches 1 while the MSE converges to 0. However, the simulations indicate that this also holds true for $\alpha = 5$. For $\alpha = 4$, it becomes questionable whether posterior contraction will eventually happen. The choice $\alpha = 3$, in contrast, leads to a clear increase of the MSE with increasing k, and posterior contraction evidently fails.

An interesting observation is the power law behavior $\sim k^{-\beta}$ of the MSE, which is revealed by linear segments in the respective log-log plots. Figure 1a shows that the slope β is independent of the chosen estimator, and 1c suggests that it might also be independent of w and μ. We can therefore consider β as a function $\beta(\alpha)$ of α alone. A numerical approximation of α^* is then given by the value of α where β changes sign, i.e.,

$$\beta(\alpha^*) = 0.$$

Since $\beta(\alpha)$ is strictly monotone, as a higher number k of observations will lead to better estimates, such an α^* is uniquely defined. Figure 2 displays an approximation of the graph of $\beta(\alpha)$ for values between $\alpha = 2$ and $\alpha = 8$. The respective slopes are estimated by linear least squares regressions for k between 10^7 and 10^9. Even though our numer-

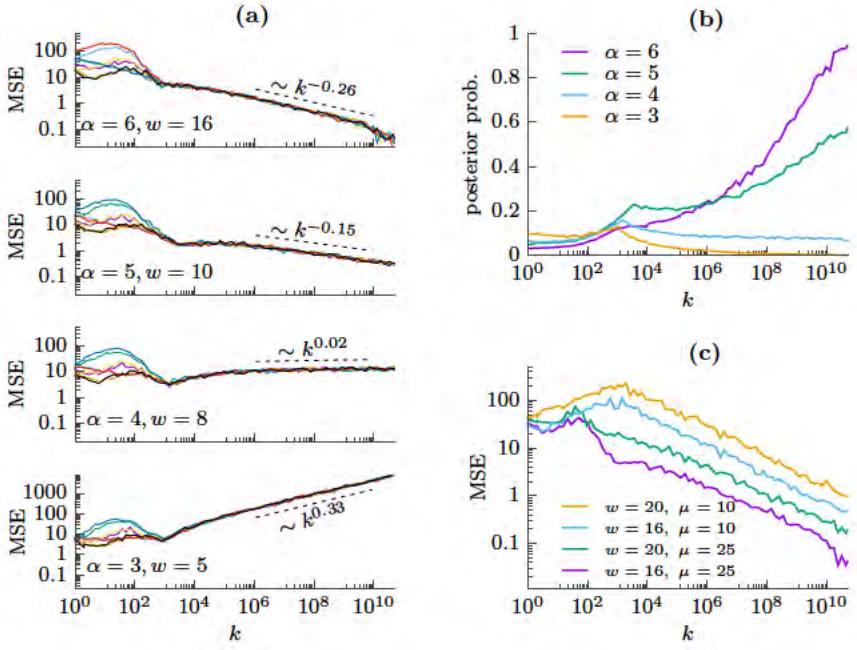


Figure 1: Asymptotic behavior of the scale estimator and posterior contraction. **(a)** shows log-log plots of the MSE of several scale estimators in different asymptotic scenarios (α, w, μ). The value μ was set to 25 in each simulation, and the parameters for the scale estimators were picked as all possible combinations of $\gamma \in \{0.5, 1\}$, $a \in \{1, 5\}$, and $b \in \{1, 5\}$. **(b)** shows the empirical mean of the posterior probabilities $\Pi(N = n_k^0 \,|\, \mathbf{X}^k)$ for the same four settings depicted in (a). **(c)** shows the MSE of the scale estimator with parameters $\gamma = a = b = 1$ for constant $\alpha = 6$ and varying values of w and μ.

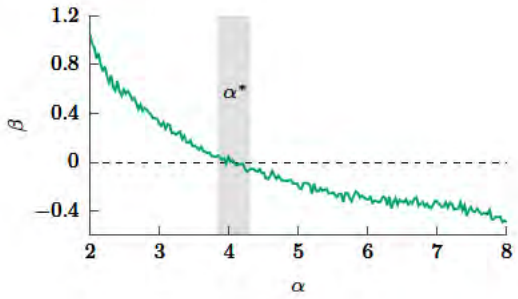


Figure 2: Relation between α and β. For a given order α, the corresponding value of β was determined by conducting simulations like in Figure 1a and fitting the slope for k between 10^7 and 10^9. The graph shows that the zero point α^* of the conjectured function $\beta(\alpha)$ has to be in the vicinity of 4.

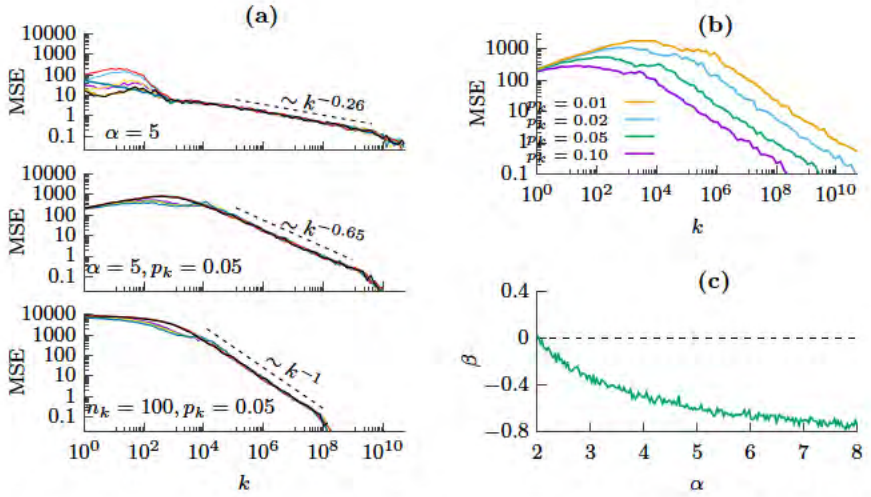


Figure 3: Comparison of alternative asymptotic settings. (a) shows the MSE for three different asymptotic scenarios. In the first plot, n_k and p_k behave like in Figure 1 with $w = 16$ and $\mu = 25$. In the second plot, p_k is fixed to the value 0.05, while n_k still increases with k ($w = 16$). The third plot addresses the scenario where both n_k and p_k are held fixed. (b) shows the scenario of growing n_k (with $\alpha = 6$ and $w = 16$) and different fixed values p_k. The graph shows that the slope in the linear segment does not depend on p_k. (c) shows the relation between β and α for the scenario with fixed p_k and growing n_k. The values of the slopes β are determined as described in Figure 2, with adapted ranges for k.

ical results do not allow us to establish the precise functional relation between α and β, it becomes clear that α^* indeed has to be close to 4.

For comparison, we additionally conducted simulations that target other asymptotic regimes. First, we keep p_k constant and let n_k again increase with the sample size, $n_k = w\sqrt[\alpha]{k}$. In this scenario, a properly rescaled binomial random variable converges to a standard normal distribution. Our simulations confirm that estimation of n is easier in this case: the MSE in Figure 3a decreases faster when $\alpha = 6$ and $p_k = 0.05$ is fixed compared to $\alpha = 6$ and $p_k \to 0$. Since the rate of convergence β in this alternative setting seems to be independent of the specific choice of $p_k = \text{const}$, see Figure 3b, we can again look at the smallest order α that still exhibits consistency. Indeed, Figure 3c reveals that the estimation of n remains consistent over a larger range of values for α in this setting, approximately as long as $\alpha > 2$ (compared to $\alpha > 4$ in the original setting).

The last asymptotic regime we consider is the classical one for parameter estimation, where $n_k = n$ and $p_k = p$ both stay constant as k grows to infinity. Figure 3a covers this regime in the last plot. It affirms that estimating n is easiest in this setting, and we obtain the expected rate $\sim k^{-1}$ for the convergence of the MSE towards zero.

4 Discussion

Theorem 1 (see [11]) shows posterior contraction under diverging parameters n_k and p_k as long as $(n_k, p_k) \in \mathcal{M}_\lambda$, which implies $n_k = O(\sqrt[6]{k/\log(k)})$. The aim of our simulation study in Section 3 was to explore the minimal rate $\sqrt[6]{k}$ for n_k such that posterior consistency remains valid. The difference in the permissible rates turns out to be rather small, since our investigation suggests that $\alpha = 5$ still allows for consistent estimation, whereas $\alpha = 3$ clearly leads to inconsistency. Figure 2 shows that the true boundary α^* is likely close to 4, indicating that Theorem 1 cannot be improved fundamentally.

Several aspects of our simulations and findings deserve further commentary. First, Figure 1c reveals that the slope β is not strongly affected by the parameters w and μ in the settings that we tested. However, our numerical approach is not suitable to verify questions like this with a high degree of confidence. For example, our numerics become instable for values $k > 10^{11}$.

Secondly, we additionally conducted simulations for other estimators than the scale estimator (2) that are not shown in the article. For example, we tested various versions of the Bayesian estimator given in [4]. While their performance for $k \leq 10^3$ varies quite much – similar to the different estimators shown in Figure 1a – their asymptotic performance is exactly the same as for the scale estimator. Notably, the maximum likelihood estimator also exhibits the very same asymptotic behavior, even though it performs poorly in the regime of smaller k. The sample maximum, in contrast, shows a completely different behavior: the MSE diverges even for $n_k \sim \log(k)$. This illustrates the sharpness of the assumptions for Lemma 10 in [11], which states that the sample maximum is consistent if $n_k \log(n_k) < c \log(k)$ for $c < 1$.

Finally, we consistently observed a phase transition in all simulations when the MSE drops below a value of about 0.1, where it changes its behavior and begins to decreases faster than $\sim k^\beta$. Indeed, it seems to decay exponentially from that point on. We conjecture that this happens due to the discreteness of n, which means that the MSE cannot measure small deviations $|\hat{n} - n| < 1$ from the real n without dropping to zero. Rather, if the posterior contracts so much that we estimate n correctly most of the time, the MSE essentially captures the probability that $\hat{n}$ lies outside of the interval $(n-1, n+1)$, and such probabilities usually decay exponentially fast. For applications, the rate of the MSE before the exponential decay is often much more interesting. One instructive example in this context is the sample maximum in the setting of fixed n and p, for which we know from Section 1 that it converges exponentially fast. However, as argued above, this only takes place when the MSE is already very small, and simulations suggest that the rate of convergence is much slower if the MSE is larger than 0.1. For instance, if $p = 0.2$ and $n = 25$, we find $\beta \approx -0.13$. Thus, even though the true asymptotic behavior of the sample maximum is exponential, the practically meaningful rate of convergence is considerably worse than the rate k^{-1} of the Bayesian estimators.

Acknowledgements

Support of the DFG RTG 2088 (B4) and DFG CRC 755 (A6) is gratefully acknowledged.

References

[1] Berger, J.O., Bernardo, J.M., Sun, D.: Objective priors for discrete parameter spaces. J. Am. Stat. Assoc. **107**, 636–648 (2012)

[2] Carroll, R.J., Lombard, F.: A note on n estimators for the binomial distribution. J. Am. Stat. Assoc. **80**, 423–426 (1985)

[3] DasGupta, A., Rubin, H.: Estimation of binomial parameters when both n, p are unknown. J. Stat. Plan. Inference **130**, 391–404 (2005)

[4] Draper, N., Guttman, I.: Bayesian estimation of the binomial parameter. Technometrics **13**, 667–673 (1971)

[5] Fisher, R.: The negative binomial distribution. Annals of Eugenics London **11**, 182–187 (1941)

[6] Günel, E., Chilko, D.: Estimation of parameter n of the binomial distribution. Commun. Stat. Simul. Comput. **18**, 537–551 (1989)

[7] Hamedani, G.G., Walker, G.G.: Bayes estimation of the binomial parameter n. Commun. Stat. Theory Methods **17**, 1829–1843 (1988)

[8] Kahn, W.D.: A cautionary note for Bayesian estimation of the binomial parameter n. Am. Stat. **41**, 38–40 (1987)

[9] Link, W.A.: A cautionary note on the discrete uniform prior for the binomial n. Ecology **94**, 2173–2179 (2013)

[10] Raftery, A.E.: Inference for the binomial n parameter: a hierachical Bayes approach. Biometrika **75**, 223–228 (1988)

[11] Schneider, L.F., Schmidt-Hieber, J., Krajina, A., Staudt, T., Aspelmeier, T., Munk, A.: Posterior consistency for n in the binomial (n,p) problem with both parameters unknown - with applications to quantitative nanoscopy. arXiv (2018)

[12] van der Vaart, A.W.: Asymptotic Statistics. Cambridge University Press (1998)

[13] Villa, C., Walker, S.G.: A cautionary note on using the scale prior for the parameter n of a binomial distribution. Ecology **95**, 2674–2677 (2014)

CHAPTER C

Threshold Selection in Univariate Extreme Value Analysis

Threshold Selection in Univariate Extreme Value Analysis

Laura Fee Schneider*[1], Andrea Krajina†[1], and Tatyana Krivobokova‡[1]

[1]Institute for Mathematical Stochastics, University of Göttingen

Abstract

Threshold selection plays a key role for various aspects of statistical inference of rare events. Most classical approaches tackling this problem for heavy-tailed distributions crucially depend on tuning parameters or critical values to be chosen by the practitioner. To simplify the use of automated, data-driven threshold selection methods, we introduce two new procedures not requiring the manual choice of any parameters. The first method measures the deviation of the log-spacings from the exponential distribution and achieves good performance in simulations for estimating high quantiles. The second approach smoothly estimates the asymptotic mean square error of the Hill estimator and performs consistently well over a wide range of distributions.
The methods are compared to existing procedures in an extensive simulation study and applied to a dataset of financial losses, where the underlying extreme value index is assumed to vary over time. This application strongly emphasizes the importance of solid automated threshold selection.

AMS 2010 Subject Classification: Primary 62G32; secondary 62G05, 62F12, 97M30

Keywords: Extreme value statistics, peak-over-threshold approach, power laws, Hill estimator, tuning parameter selection, bias estimation

*laura-fee.schneider@mathematik.uni-goettingen.de
†akrajina@gmail.com
‡tkrivob@gwdg.de

1 Introduction

Extreme value analysis of heavy-tailed distributions is an important model in various applications. In seismology and climatology, for example, statistics of extremes is used to study earthquakes (Beirlant et al., 2018) or heavy precipitation (Carreau et al., 2017). Another important field of research is analysing high financial losses, which becomes particularly interesting if the losses depend on covariates (Chavez-Demoulin et al., 2016; Hambuckers et al., 2018). In this situation an automated threshold selection procedure could bring additional benefits by enabling the selection of the threshold depending on a covariate. We will discuss this possibility in more detail in Section 5.

To mathematically investigate the behaviour of heavy tails, we consider random variables from the domain of attraction (DoA) of a Fréchet distribution. Let $X_1, \dots, X_n$ be independent identically distributed (i.i.d.) random variables with distribution function F, where F is in the DoA of an extreme value distribution (evd) G_γ with extreme value index $\gamma > 0$. This means there exist sequences $a_n > 0$ and b_n real, s.t.

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G_\gamma(x) := \exp\left(-x^{-1/\gamma}\right).$$

In this situation the following first order condition holds,

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1/\gamma}, \tag{1}$$

i.e. the survival function $1 - F$ is regularly varying with index $-1/\gamma$. Distributions fulfilling this condition are called Pareto-type distributions, because they only differ from the Pareto distribution by a slowly varying function $\ell_F(x)$, i.e. $1 - F(x) = x^{-1/\gamma}\ell_F(x)$.

We can interpret the quotient in (1) as a conditional probability, and it follows directly that

$$\begin{aligned} \frac{X_1}{t} \,\Big|\, X_1 > t \ &\xrightarrow{\mathcal{D}} P, \text{ as } t \to \infty \text{ and } P \sim \text{Pareto}\left(1, \frac{1}{\gamma}\right), \\ \log\left(\frac{X_1}{t}\right) \,\Big|\, X_1 > t \ &\xrightarrow{\mathcal{D}} E, \text{ as } t \to \infty \text{ and } E \sim \text{Exp}\left(\frac{1}{\gamma}\right). \end{aligned} \tag{2}$$

Thus, for a sufficiently large threshold t the data above this threshold can be modelled by a Pareto or an exponential distribution. In this article we concentrate on the exponential approximation and utilize it for inference on

the extreme value index. It is common to consider the threshold $t = X_{(n-k,n)}$ and choose the sample fraction k instead of t, where $X_{(1,n)} \leq \cdots \leq X_{(n,n)}$ denote the order statistics of a sample of size n. In this case, a natural estimator for γ under the exponential approximation of the log-spacings $Y_{(i,k)} := \log(X_{(n-i+1,n)}) - \log(X_{(n-k,n)})$ is their mean, the Hill estimator (Hill, 1975),

$$\hat{\gamma}_k := \frac{1}{k} \sum_{i=1}^{k} \log \left(\frac{X_{(n-i+1,n)}}{X_{(n-k,n)}} \right) = \frac{1}{k} \sum_{i=1}^{k} Y_{(i,k)}. \tag{3}$$

The Hill estimator is still among the most popular and well-known estimators for the extreme value index, although its sample path as a function in k can be highly unstable and estimation therefore crucially depends on the choice of the sample fraction k. This dependence highlights the difficulties in estimating γ: even from univariate i.i.d. observations from $F \in \mathrm{DoA}(G_\gamma)$, estimation is hard, since only few observations contain information about the extreme value distribution G_γ. To select a threshold above which the data can be used for statistical inference about the tail is one of the most fundamental problems in the field of extreme value analysis.

Due to the importance of this task, the appropriate choice of the threshold has been discussed extensively in extreme value research over the last decades, and suggested solutions cover a variety of methodologies. We give a short summary on different types of approaches and stress the specific difficulties that arise. We mainly concentrate on methods we compare in our simulation study in Section 4. More comprehensive reviews about threshold selection can be found in Scarrott and MacDonald (2012) and Dey and Yan (2016).

One basic concept in threshold selection is data visualisation, which is also discussed more deeply in Kratz and Resnick (1996) and Drees et al. (2000). Popular graphical diagnostics used in this context are the Zipf plot, Hill plot, QQ-plot or the mean-excess plot to name a few. A major drawback of these methods is their subjectivity due to the necessarily personal interpretation of the plot. Further, it is a burden to choose each threshold manually, especially in high dimensional settings or when analysing many samples. Easier ways to select the sample fraction are rules-of-thumb such as using the upper 10% of the data (DuMouchel, 1983) or $k = \sqrt{n}$ (Ferreira et al., 2003). However, these suggestions are neither theoretically justified nor data driven. Reiss and Thomas (2007) present a procedure that tries to find a region of stability among the estimates of the extreme value in-

dex. Their method depends on a tuning parameter, whose choice is further analysed in Neves and Fraga Alves (2004). To our knowledge no theoretical analysis exists for this approach.
Besides these and similar heuristic approaches, there is a class of theoretically motivated procedures that target the optimal sample fraction for specific estimation tasks, such as quantile estimates (Ferreira et al., 2003), estimation of high probabilities (Hall and Weissman, 1997) or the Hill estimator, see below. We also mention two other methodologies. First, there are suggestions that utilize comparing the empirical distribution to the fitted generalized Pareto distribution (GPD) via goodness-of-fit tests (Bader et al., 2018) or by minimizing the distance between them (Pickands, 1975; Gonzalo and Olmo, 2004; Clauset et al., 2009), where the latter approach is theoretically analysed by Drees et al. (2018). Further, Goegebeur et al. (2008) propose a family of kernel statistics to test for exponentiality in order to select a threshold.
Of particular interest to us are methods that aim to estimate the sample fraction k_{opt} which minimizes the asymptotic mean square error (AMSE) of the Hill estimator. To construct an estimator for k_{opt}, Drees and Kaufmann (1998) utilize the Lepskii method and an upper bound on the maximum random fluctuation of $\hat{\gamma}_k$ around γ. To apply their approach it is necessary to choose several tuning parameters and to obtain consistent initial estimates for γ and a second order parameter ρ. They recommend specific choices of the parameters based on a numerical study and we employ their proposals in our simulations. However, the choice of these parameters is not data-driven. In Guillou and Hall (2001), a test statistic Q_k is constructed based on an accumulation of log-spacings, which takes values around 1 as long as the bias of the Hill estimator is not significantly large. Their statistic depends on a tuning parameter as well, and a critical value to test Q_k against has to be chosen. Again we adopt the parameter choice suggested in their simulation study. Danielsson et al. (2001) introduce a double bootstrap approach to estimate the optimal sample fraction. They need to choose the number of bootstrap samples and a parameter n_1. For n_1, a data-driven but computationally expensive selection method is provided, where the whole bootstrap procedure is repeated for various possible values of n_1. Another estimator for k_{opt} is given by Beirlant et al. (2002), which employs least squares estimates from an exponential regression approach. The method depends on an estimate for ρ and a sample fraction k_0. To avoid the choice of k_0 they suggest taking the median of the estimates over a range of values, e.g. $k_0 \in \{3, \dots, n/2\}$. A different approach is taken by Goegebeur et al. (2008), who use the properties of a test statistic regarding bias estimation

to construct an estimator for the AMSE/γ and minimize it with respect to k. If one fixes $\rho = -1$, as they suggest in their simulations chapter, there is no further tuning parameter to be chosen. However, no result about consistency of $\hat{k}$ in the sense of $\hat{k}/k_{\text{opt}} \xrightarrow{\mathbb{P}} 1$ is known in contrast to the approaches in Drees and Kaufmann (1998), Guillou and Hall (2001), Danielsson et al. (2001) and Beirlant et al. (2002).

In this paper we contribute to the problem of threshold selection by introducing two new methods. The first one presented in Section 2 is inspired by the idea of testing the exponential approximation. We estimate the integrated square error (ISE) of the exponential density under the assumption that the log-spacings are indeed exponentially distributed. The error functional we obtain, denoted as inverse Hill statistic (IHS), is very easy to compute and does not depend on any tuning parameters. Since this criterion is variable for small k, it can be additionally smoothed to improve the performance. The minimizing sample fraction of IHS is asymptotically smaller than k_{opt}, as it is stricter against deviation from the exponential approximation. This estimator performs remarkably well for adaptive quantile estimation on finite samples, as illustrated in our simulation study.
In our second approach we suggest a smooth estimator for the AMSE of the Hill estimator, called SAMSEE (smooth AMSE estimator). This estimator is constructed by a preliminary estimate of γ using the generalized Jackknife approach in Gomes et al. (2000) and a bias estimator for the Hill estimator introduced in Section 3. By minimizing SAMSEE we estimate the optimal sample fraction k_{opt}. For estimation, the choice of a large sample fraction K is necessary, for which we present a data-driven selection procedure in Section 3. SAMSEE utilizes the idea of fixing $\rho = -1$, which is justified by good performance in simulations and leads to a simpler and more robust estimator. However, the estimator can also be adjusted to any ρ by including a consistent estimator $\hat{\rho}$, as described in Section 3.1.
After introducing our two novel threshold selection methods in Sections 2 and 3 we compare these methods to various other approaches in an numerical analysis in Section 4. In Section 5 the importance of automated threshold selection procedures is illustrated in an application, where we nonparametrically estimate an extreme value index that varies over time. The proof of Theorem 3, which describes the asymptotic behaviour of our bias estimator, and auxiliary theoretical results can be found in Appendix A.

2 IHS – The inverse Hill statistic

In this section we introduce the first threshold selection procedure by analysing the integrated square error (ISE) between the exponential density h_γ and its parametric estimator $h_{\hat{\gamma}_k}$ employing the Hill estimator,

$$\mathrm{ISE}(k) := \int \left(h_\gamma(x) - h_{\hat{\gamma}_k}(x)\right)^2 \mathrm{d}x = \frac{1}{2\gamma} - \frac{2}{\gamma + \hat{\gamma}_k} + \frac{1}{2\hat{\gamma}_k}.$$

The first term of ISE is constant and thus plays no role for selecting k. The last term of ISE is known, but the second term is not. Therefore, we cannot minimize ISE directly. Instead, we want to estimate and minimize its expectation under the exponential approximation. This is based on the idea of considering the hypothesis H_0 that the log-spacings $Y_{(i,k)}$ are indeed exponentially distributed. Under H_0 the Hill estimator is gamma distributed, see Lemma 1, and the mean of ISE (MISE) can be calculated explicitly. We observe that MISE is a decreasing function in k under the exponential approximation,

$$\mathrm{MISE}(k) - \frac{1}{2\gamma} := \mathbb{E}_{H_0}[\mathrm{ISE}(k)] - \frac{1}{2\gamma} = -\frac{1}{\gamma}C(k) + \frac{k}{2(k-1)\gamma}, \qquad (4)$$

where $C(k) := 2\exp(k)k^k\Gamma(1-k,k)$ and $\Gamma(a,b)$ denotes the upper incomplete gamma function. The function $C(k)$ converges to 1 very fast, s.t. we obtain

$$\mathbb{E}_{H_0}\left[\frac{2}{\gamma + \hat{\gamma}_k}\right] \approx \frac{1}{\gamma} = \mathbb{E}_{H_0}\left[\frac{k-1}{k\hat{\gamma}_k}\right].$$

This provides us with an unbiased estimator for the first term in (4) under H_0. However, due to the high variability for small k, we instead want to find an estimator of the form $w/\hat{\gamma}_k$ for some w depending on k that minimizes the MSE under the exponential approximation. To do so, we approximate its MSE in the following way,

$$\mathbb{E}_{H_0}\left[\left(\frac{w}{\hat{\gamma}_k} - \frac{2}{\hat{\gamma}_k + \gamma}\right)^2\right] \approx \frac{w^2k^2}{\gamma^2(k-1)(k-2)} - \frac{2wk}{\gamma^2(k-1)} + \frac{1}{\gamma^2}. \qquad (5)$$

The approximation depends on similar functions as $C(k)$, which quickly become constant. The MSE in (5) is minimized for $w = (k-2)/k$. Thus, we suggest the inverse Hill statistic

$$\mathrm{IHS}(k) := \frac{1}{2\hat{\gamma}_k} - \frac{k-2}{\hat{\gamma}_k k} = \frac{4-k}{2\hat{\gamma}_k k}$$

to estimate $\text{MISE}(k) - (2\gamma)^{-1}$ and the threshold selected via minimizing IHS,

$$\hat{k}_{\text{IHS}} := \underset{1<k<n}{\arg\min}\ \text{IHS}(k).$$

By minimizing IHS we select a sample fraction where IHS starts increasing and contradicts H_0 by behaving contrarily to MISE under the exponential approximation. This criterion can be compared to hypothesis testing with a large significance level α, which implies seeking high confidence when deciding to not reject H_0. Further properties of $\hat{k}_{\text{IHS}}$ are analysed theoretically in Section 2.1 and for finite samples in a numerical study in Section 4.
Note that the performance of IHS depends on the bias of the Hill estimator being positive and increasing, see Section 2.1. However, the bias can be negative for some non-standard distributions. In case of a negative bias, we instead suggest to use,

$$\text{IHS}^{-}(k) := \frac{4+k}{2\hat{\gamma}_k k} \quad \text{and} \quad \hat{k}_{\text{IHS}^-} := \underset{1<k<n}{\arg\min}\ \text{IHS}^{-}(k).$$

The two cases can easily be distinguished by analysis of the Hill estimator for large k. Both IHS and IHS^{-} are justified by asymptotic results in Section 2.1.
Figure 1 illustrates that IHS is highly varying for small k, which makes automatic threshold choices more variable. To control this problematic behaviour we smooth the IHS. More specifically, we want to estimate $\mathbb{E}[\text{IHS}]$ by considering the regression problem

$$\text{IHS}(k) = \mathbb{E}[\text{IHS}](k) + \sigma\epsilon_k, \quad k = 1, \ldots, n,$$

where $\sigma > 0$ and $\mathbb{E}[\epsilon_k] = 0$. Due to the structure of the Hill estimator, the random variables ϵ_k are highly dependent, which needs to be taken into account in estimation. In our simulations, we apply a Bayesian nonparametric procedure introduced by Serra et al. (2018) which simultaneously estimates mean and covariance and is available in the R-package *eBsc*. The approach provides a smooth estimator for the expectation of IHS – denoted as sIHS – comprising less variation for small k. This way we can improve the performance by selecting a more suitable threshold, as illustrated in Figure 1. Of course, one can also use other smoothing procedures suitable for dependent data (Opsomer et al., 2001; Krivobokova and Kauermann, 2007; Lee et al., 2010).

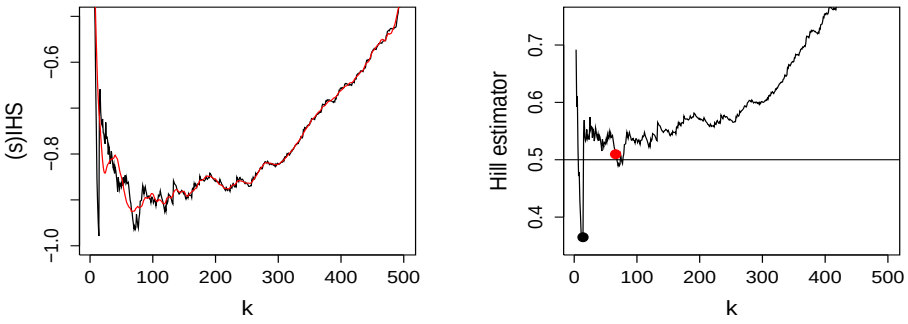


Figure 1: On the left, the IHS (dashed) and sIHS (red) are plotted for a Fréchet(2) sample of size $n = 500$. On the right the Hill plot for the same sample with the minimizing k of IHS (black) and sIHS (red) is shown, where the dotted line marks the true value of $\gamma = 1/2$.

We finally want to remark on the relation between IHS and ISE, which is given by

$$\text{IHS} + \frac{1}{2\gamma} = \text{ISE} + \frac{2}{k\hat{\gamma}_k} + \frac{\hat{\gamma}_k - \gamma}{\hat{\gamma}_k(\hat{\gamma}_k + \gamma)}. \tag{6}$$

This equation points out that minimizing IHS does not minimize ISE, as IHS takes an additional bias term into account. If the bias of the Hill estimator is positive, IHS selects smaller k (larger thresholds) than ISE. This is not surprising, because we estimate the expectation of the ISE under the hypothesis that the exponential approximation holds. This is a much more conservative error functional, meaning it is more strict against deviation from the exponential distribution.
In conclusion, with IHS we do not aim to estimate k_{opt} but to find a sample fraction where we can be very certain that the exponential approximation still holds. The impact of this consideration is illustrated in simulations and an application in Sections 4 and 5.

2.1 Theorectical analysis of IHS

In order to understand the IHS asymptotically we consider the second order condition,

$$\lim_{t\to\infty} \frac{\frac{U(tx)}{U(t)} - x^\gamma}{A(t)} = x^\gamma \frac{x^\rho - 1}{\rho}, \tag{7}$$

for $x > 0$ and with second order parameter $\rho < 0$. Here, $A(t)$ denotes a function converging to zero as t goes to infinity and $|A|$ is regularly varying with index ρ. Further, U is defined by $U(x) := F^{\leftarrow}\big(1 - \frac{1}{x}\big)$, where $F^{\leftarrow}$ denotes the left inverse of the distribution function F. In this setting the following asymptotic normality statements for the Hill estimator $\hat{\gamma}_k$ hold.

Theorem 1 (Theorem 3.2.5 in de Haan and Ferreira (2006))**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution function $F \in \mathrm{DoA}(G_\gamma)$ for $\gamma > 0$. If (7) holds and k is an intermediate sequence, i.e. $k \to \infty$ and $k/n \to 0$ as $n \to \infty$, then*

$$\sqrt{k}(\hat{\gamma}_k - \gamma) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\frac{\lambda}{(1-\rho)}, \gamma^2\right),$$

with $\lambda := \lim\limits_{k\to\infty} \sqrt{k} A(n/k)$.

Theorem 2. *Under the conditions of Theorem 1, it holds that*

$$\sqrt{k}\left(\frac{1}{\hat{\gamma}_k} - \frac{1}{\gamma}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\frac{-\lambda}{(1-\rho)\gamma^2}, \frac{1}{\gamma^2}\right).$$

Proof. Applying the delta method to Thm. 1. □

Following the reasoning in de Haan and Ferreira (2006), page 78, the minimizing point of the AMSE can be found explicitly if considering $A(t) = ct^\rho$ with $c \neq 0$. In this special case the minimizing sample fraction can be expressed as

$$k_{\mathrm{opt}} = \left\lfloor \left(\frac{\gamma^2(1-\rho)^2}{-2\rho c^2}\right)^{1/(1-2\rho)} n^{-2\rho/(1-2\rho)} \right\rfloor. \tag{8}$$

Under the same assumption we can calculate the minimizing point k_{IHS} of the asymptotic expectations of IHS and IHS$^-$. Let $\mathbb{AE}$ denote the asymptotic expectation referring to the expectation of the limiting distribution in Thm. 2. Then

$$\begin{aligned} k_{\mathrm{IHS}} &:= \arg\min_k \ \mathbb{AE}[\mathrm{IHS}] = \arg\min_k \left\{ \frac{2}{\gamma k} + \frac{A(n/k)}{2\gamma^2(1-\rho)} \cdot \frac{k-4}{k} \right\} \\ &\approx \arg\min_k \left\{ \frac{2}{\gamma k} + \frac{A(n/k)}{2\gamma^2(1-\rho)} \right\} = \left\lfloor \left(\frac{4\gamma(1-\rho)}{-\rho c}\right)^{1/(1-\rho)} n^{-\rho/(1-\rho)} \right\rfloor. \end{aligned}$$

It is easy to check that the same formula holds for IHS$^-$ if c is replaced by its absolute value. Further note that by Lemma 2 it is sufficient to

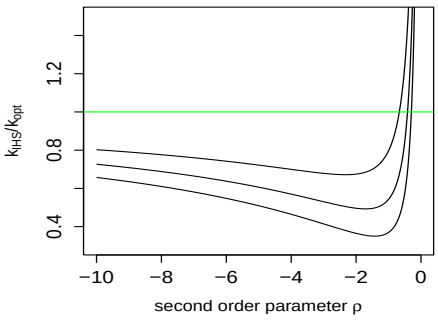


Figure 2: The approximation of the proportion $k_{\text{IHS}}/k_{\text{opt}}$ in (9) is plotted as a function in ρ for $\gamma = c = 1$ and $n = 500$ (solid), $n = 5000$ (dashed) and $n = 50000$ (dotted).

consider intermediate sequences when determining the minimizing sequence. Comparing k_{opt} and k_{IHS} for a fixed $\rho > -\infty$ we obtain that

$$\frac{k_{\text{IHS}}}{k_{\text{opt}}} \approx \left(\frac{-\rho}{32} \cdot k_{\text{IHS}}\right)^{-1/(1-2\rho)} \approx d \cdot n^{\frac{\rho}{(1-2\rho)(1-\rho)}} \longrightarrow 0, \tag{9}$$

as $n \to \infty$ and for a constant d depending on ρ, γ and c. This supports what equation (6) already suggested: minimizing IHS gives asymptotically a smaller k than k_{opt}. Thus, k_{IHS} asymptotically performs suboptimally for the Hill estimator but still leads to a consistent sequence of estimates. For finite samples the ratio crucially depends on ρ, and k_{IHS} can be even larger than k_{opt}, as illustrated in Figure 2. The graphic presents the quotient of the two sample fractions as a function in the second order parameter ρ for different samples sizes. The parameters c and γ are fixed to 1, as they have a weaker impact on the proportion. It also holds that $k_{\text{IHS}}/k_{\text{opt}} \to 1$, as $\rho \to -\infty$, since both sample fractions converge to n in this case.

Although k_{IHS} is of smaller order than k_{opt} asymptotically, the simulation study in Section 4 shows that $\hat{k}_{\text{IHS}}$ works remarkably well when used for quantile estimation. We consider the following quantile estimator for the $(1-p)$-quantile,

$$\hat{q}_k(p) = X_{(n-k,n)} \left(\frac{k}{np}\right)^{\hat{\gamma}_k}. \tag{10}$$

The sample fraction k_{opt} also minimizes the asymptotic relative MSE of $\hat{q}_k(p)$, see e.g. Theorem 4.3.8 in de Haan and Ferreira (2006). For finite samples however, the quantile estimator seems to benefit from k_{IHS}. This has different reasons, two of which are illustrated by Figure 3. On the left

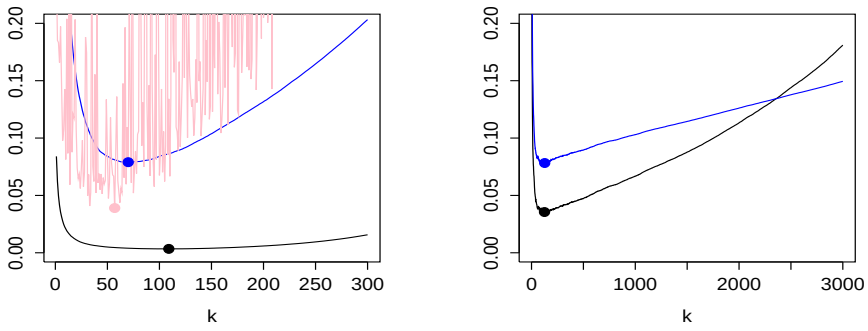


Figure 3: Empirical expectations of IHS (blue), MSE (black) and MSEQ (pink). The left plot is based on 10,000 samples from a Fréchet(2) distribution of size 500. The graphic on the right is based on 500 samples of size 5000 from a Loggamma distribution.

we see the empirical expectation of IHS, the empirical versions of the MSE of $\hat{\gamma}_k$ and the relative MSE of the quantile estimator,

$$\text{MSEQ} := \mathbb{E}\left[\left(\frac{\hat{q}_k(p)}{q(p)} - 1\right)^2\right] / \log\left(\frac{k}{np}\right), \tag{11}$$

as used in Theorem 4.3.8 in de Haan and Ferreira (2006). We observe that k_{IHS} (blue dot) is indeed smaller than k_{opt} (black) but so is the minimizer of MSEQ (pink) as well.
On the right we see a plot of the empirical $\mathbb{E}$[IHS] and MSE of $\hat{\gamma}_k$ for Loggamma distributed samples of size 5000. This graphic highlights the similarities between MSE and IHS for the boundary case $\rho = 0$.
These observations indicate why $\hat{k}_{\text{IHS}}$ outperforms other methods that try to minimize the MSE of the Hill estimator when adaptively estimating $q(p)$ by (10) on most of our exemplary distributions and sample sizes $n = 500$ and $n = 5000$, see Section 4.

3 SAMSEE - The smooth AMSE estimator

In this section we illustrate a way to smoothly estimate the AMSE of the Hill estimator. Via minimizing this AMSE estimator, called SAMSEE, we obtain an estimator for k_{opt}. By this means, we extend previous methods

which also estimate k_{opt} by estimating the AMSE itself. From Thm. 1 it is easy to see that the AMSE, which is the asymptotic variance plus the asymptotic squared bias, equals

$$\mathbb{AE}[(\hat{\gamma}_k - \gamma)^2] = \frac{\gamma^2}{k} + \frac{A(n/k)^2}{(1-\rho)^2}. \tag{12}$$

Thus, to estimate the AMSE as a function in k we employ two estimators, one for γ and one for the bias term as a combination of ρ and A. First we explain how we estimate γ and then we define the bias estimator. This bias estimator has a quite smooth sample path in k, and it depends on the choice of a large sample fraction K, for which we afterwards provide a data-driven selection procedure.

Note that, for the moment, we assume that the second order parameter ρ is equal to -1 to motivate the construction of the AMSE estimator. The idea of misspecifying ρ to simplify estimation – via avoiding the additional uncertainty through estimating ρ or selecting an influential tuning parameter – was already used, for example, by Gomes et al. (2000), Drees and Kaufmann (1998) and Goegebeur et al. (2008). It is also motivated by the simulations in Section 3.1.

For γ we consider the generalized Jackknife estimator $\hat{\gamma}_k^{\text{GJ}}$ introduced by Gomes et al. (2000) as $\gamma_n^{\text{G}_1}$. This estimator is defined by

$$M_{n,k} := \frac{1}{k}\sum_{i=1}^{k} Y_{(i,k)}^2, \quad \hat{\gamma}_{\text{V},k} := \frac{M_{n,k}}{2\hat{\gamma}_k}, \text{ and } \quad \hat{\gamma}_k^{\text{GJ}} := 2\hat{\gamma}_{\text{V},k} - \hat{\gamma}_k, \tag{13}$$

where $Y_{(i,k)}$ denotes the log-spacings as in equation (3). Note, that $\hat{\gamma}_{V,k}$ is the de Vries estimator introduced under this name in de Haan and Peng (1998) and $\hat{\gamma}_k$ is the Hill estimator as above. The generalized Jackknife estimator has a reduced bias compared to the Hill estimator and is even asymptotically unbiased if $\rho = -1$, see (2.11) in Gomes et al. (2000). This property is useful here, since the bias estimator $\bar{b}_{\text{up},K,k}$ defined in the following performs optimally for $\rho = -1$ as well. Furthermore, the same large sample fraction K can be used for $\hat{\gamma}_K^{\text{GJ}}$ and $\bar{b}_{\text{up},K,k}$.

To construct this bias estimator, we study the following averages of Hill estimators,

$$\bar{\gamma}_k := \frac{1}{k}\sum_{i=1}^{k} \hat{\gamma}_i \quad \text{and} \quad \bar{\gamma}_{\text{up},K,k} := \frac{1}{K-k+1}\sum_{i=k}^{K} \hat{\gamma}_i,$$

where $k < K$. Plotting these averages illustrates how they smoothly frame the sample path of the Hill estimator. Especially the upper mean $\bar{\gamma}_{\text{up},K,k}$

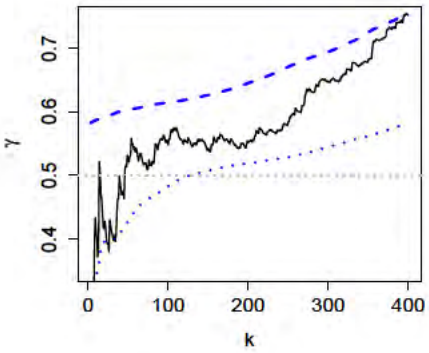


Figure 4: The plot shows the Hill estimator (black), $\bar{\gamma}_k$ (blue dotted) and $\bar{\gamma}_{\mathrm{up},K,k}$ (blue dashed) with $K = 400$ for a Fréchet(2) sample of size $n = 500$ with true extreme value index $1/2$.

seems to contain a lot of structural information about the underlying asymptotic bias of the Hill estimator when choosing the upper bound K appropriately, see Figure 4. This similarity between the upper mean and the bias of the Hill estimator inspires the definition

$$\bar{b}_{\mathrm{up},K,k} := \bar{\gamma}_{\mathrm{up},K,k} - \bar{\gamma}_K. \tag{14}$$

The estimator $\bar{b}_{\mathrm{up},K,k}$ is indeed a sensible estimator for a bias function, since

$$\mathbb{AE}[\bar{b}_{\mathrm{up},K,k}] = \frac{-\rho A(n/k)}{(1-\rho)^2} = \frac{1}{2}\frac{A(n/k)}{(1-\rho)} \tag{15}$$

follows for $\rho = -1$ from Theorem 3.
Danielsson et al. (2001) use $(\hat{\gamma}_{\mathrm{V},k} - \hat{\gamma}_k)$ to access the bias of $\hat{\gamma}_k$ and apply a double bootstrap procedure to stabilize this highly varying estimate. We use the difference of two estimators for γ as well, but now consider averaging to smooth the bias estimate. The idea to average the Hill estimator in order to smooth the Hill plot and decrease the variance is also studied in Resnick and Stărică (1997).
It remains to choose an appropriate K in order to complete SAMSEE and to estimate the optimal sample fraction k_{opt}. We need K to be large enough to allow for minimization over all relevant k and small enough to be an intermediate sequence itself (see Theorem 3 for this condition). To find such a K we use the following relation between the estimators,

$$\mathbb{AE}[\hat{\gamma}_k] = \mathbb{AE}[\hat{\gamma}_{\mathrm{V},k} + \bar{b}_{\mathrm{up},K,k}]. \tag{16}$$

This provides us with a relatively stable function in k, $\hat{\gamma}_{\mathrm{V},k} + \bar{b}_{up,K,k}$, that has the same asymptotic expectation as the highly non-smooth Hill estimator.

We want to find an intermediate sequence K for which (16) holds and thus define

$$E^2(K) := \frac{1}{K} \sum_{k=1}^{K} \left(\hat{\gamma}_{\mathrm{V},k} + \bar{b}_{\mathrm{up},K,k} - \hat{\gamma}_k \right)^2 \tag{17}$$

to measure the deviation from approximation (16) uniformly over all $k \leq K$. Based on this, we suggest to choose

$$K^* := \arg\min_K \left\{ \sum_{L=K-2}^{K+2} \left(E^2(K) - E^2(L) \right)^2 \right\}. \tag{18}$$

In this way we select a K^* where the asymptotic approximation (16) is most stable, since we minimize the local variation of $E^2(K)$. Simulations suggest that this criterion is not sensitive to slightly increasing the region of stability $\{K-h, \ldots, K+h\}$ from $h = 2$ to $h = 5$ or 10 depending on the sample size.

Now we finally combine the previously described estimators to approach the AMSE in (12) under the assumption that $\rho = -1$. With K^* in (18) and the property of $\bar{b}_{\mathrm{up},K,k}$ in (15), we obtain an estimator for the AMSE of the Hill estimator and for k_{opt} by

$$\begin{aligned} \mathrm{SAMSEE}(k) &:= \frac{(\hat{\gamma}^{\mathrm{GJ}}_{K^*})^2}{k} + 4\bar{b}^2_{\mathrm{up},K^*,k}, \\ \hat{k}_{\mathrm{SAMSEE}} &:= \operatorname*{argmin}_{1<k<K^*} \mathrm{SAMSEE}(k). \end{aligned} \tag{19}$$

Figure 5 illustrates how such a smooth estimate of the AMSE can look like. On the left, SAMSEE is displayed for a Fréchet sample with parameters $\gamma = 1/2$ and $\rho = -1$. On the right, the Hill plot of the same sample is presented for all $k \leq K^* = 388$.

This smooth estimate of the AMSE can be useful beyond the context of threshold selection. For extreme value mixture models or Bayesian threshold selection approaches, SAMSEE could be used to construct a transition function between bulk and tail distribution or an empirical prior for the threshold, respectively, see Scarrott and MacDonald (2012) for a review on mixture models.

3.1 SAMSEE if $\rho \neq -1$

We next want to analyse SAMSEE in the broader context of an unknown second order parameter ρ. The first thing to note is that the generalized

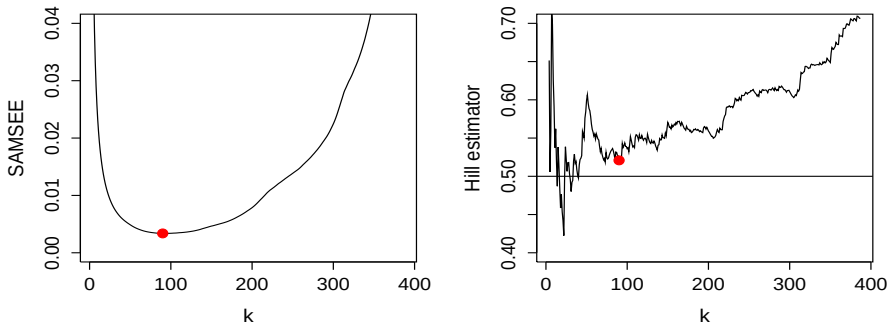


Figure 5: SAMSEE with $K^* = 388$ on the left next to the Hill plot for the same Fréchet(2) random sample of size $n = 500$ for $k \leq K^*$. The red dot indicates the selected sample fraction $\hat{k}_{\text{SAMSEE}}$ and the adaptive Hill estimate $\hat{\gamma}_{\hat{k}_{\text{SAMSEE}}}$ in the right plot.

Jackknife estimator is no longer unbiased in this situation. Secondly, the behaviour of our bias estimator changes, as it is described in the following Theorem.

Theorem 3. *Under the conditions of Theorem 1 and for $k/K \to c$ with $0 < c < 1$ as $n \to \infty$, it holds for $\bar{b}_{\text{up},K,k}$ in (14) that*

$$\sqrt{k} \cdot \bar{b}_{\text{up},K,k} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\frac{-\rho\lambda}{(1-\rho)^2} \delta_\rho(c),\ \gamma^2 \nu(c) \right),$$

where $\delta_\rho(c) = (c^\rho - 1)/(-\rho(c^{-1} - 1))$ and $\nu(c) = 2c^2/(1-c)^2 \cdot (1 - c + c\log(c))$ with $0 \leq \nu(c) \leq 1$.

Proof. The proof can be found at the end of Appendix A. □

From Theorem 3 follows that

$$\mathbb{AE}[\bar{b}_{\text{up},K,k}] = \frac{-\rho A(n/k)}{(1-\rho)^2} \cdot \delta_\rho(k/K).$$

For $\rho = -1$ the function $\delta_{-1}(c)$ is equal to 1. If $\rho \neq -1$, we can observe that δ bends our bias estimator and it will therefore increase slightly too fast or too slow. We can still apply SAMSEE in this situation and select K^* from

(18). However, approximation (16) does not hold anymore and instead the following holds,

$$\mathbb{AE}[\hat{\gamma}_k - \hat{\gamma}_{\mathrm{V},k}] - \mathbb{AE}[\bar{b}_{\mathrm{up},K,k}] = \frac{-\rho A(n/k)}{(1-\rho)^2}\left(1 - \delta_\rho(k/K)\right). \tag{20}$$

The absolute value of the error described by (20) is high if δ_ρ strongly differs from 1 and the bias term A is large. If $\rho \neq -1$, δ_ρ indeed deviates from 1 and we minimize the error by minimizing the bias. This is why applying (18) in this case leads to a small K^*. On the other hand, if $\delta_\rho = 1$, the approximation stays valid for an increasing bias and K^* will typically be larger.

An alternative to fixing $\rho = -1$ is to incorporate a consistent estimator $\hat{\rho}$ of the second order parameter. This can be done via

$$K^*_{\hat{\rho}} := \operatorname*{arg\,min}_K \left\{ \sum_{L=K-2}^{K+2} \left(E^2_{\hat{\rho}}(K) - E^2_{\hat{\rho}}(L) \right)^2 \right\},$$
$$\text{where } E^2_{\hat{\rho}}(K) := \frac{1}{K}\sum_{k=1}^{K} \left(\hat{\gamma}_{\mathrm{V},k} + \bar{b}_{\mathrm{up},K,k}/\delta_{\hat{\rho}}(k/K) - \hat{\gamma}_k \right)^2.$$

and

$$\mathrm{SAMSEE}_{\hat{\rho}}(k) := \frac{(\hat{\gamma}^{\mathrm{GJ}}_{K^*_{\hat{\rho}}})^2}{k} + \left((1-\hat{\rho}) \frac{K^*_{\hat{\rho}}/k - 1}{(k/K^*_{\hat{\rho}})^{\hat{\rho}} - 1} \cdot \bar{b}_{\mathrm{up},K^*_{\hat{\rho}},k} \right)^2, \tag{21}$$
$$\hat{k}_{\hat{\rho},\mathrm{SAMSEE}} := \operatorname*{argmin}_{1<k<K^*} \mathrm{SAMSEE}_{\hat{\rho}}(k).$$

In this way we can construct an estimator for k_{opt} in the general setting of Pareto-type distributions.
In Table 1, we present the results of a simulation study indicating for which distributions it is beneficial to use $\hat{\rho}$ instead of $\rho = -1$. We estimate ρ using the estimator $\hat{\rho}^{(1)}$ suggested in Theorem 1 in Drees and Kaufmann (1998). The results indicate that, in general, it is sensible to fix $\rho = -1$ in SAMSEE, since only for the Cauchy distribution using $\hat{\rho}$ performs slightly better regarding bias and RMSE. This confirms the observations already made by others (Gomes et al., 2000; Drees and Kaufmann, 1998; Goegebeur et al., 2008), that it is often recommendable to select $\rho = -1$ instead of allowing for further variability by including an additional estimator.

	γ	ρ	$\mathbb{E}[\hat{\gamma}_{\hat{k}}]$ (RMSE) true ρ	$\rho = -1$	$\hat{\rho}$
Student-t(6)	0.17	-1/3	0.21 (0.09)	0.26 (0.12)	0.28 (0.14)
Fréchet(2)	0.50	-1	0.51 (0.07)	0.51 (0.07)	0.51 (0.08)
Cauchy	1.00	-2	1.01 (0.13)	0.97 (0.17)	0.99 (0.16)
Burr(2,1)	2.00	-1	2.05 (0.34)	2.05 (0.34)	2.03 (0.40)

Table 1: The averages of adaptive γ estimates and their root mean square error (RMSE) in brackets are presented for thresholds $\hat{k}$ that are selected using SAMSEE or SAMSEE$_{\hat{\rho}}$ with the true ρ, $\rho = -1$ or $\hat{\rho} = \hat{\rho}^{(1)}$.

4 Simulation study

In the following we numerically analyse the performance of eight threshold selection methods on heavy-tailed distributions with very different tail behaviour. The simulation study is based on the following distributions:

- the Student-t distribution with 6 degrees of freedom, which corresponds to $\gamma = 1/6$ and $\rho = -1/3$,
- the Fréchet distribution with parameter $\alpha = 2$ and distribution function $F(x) = \exp(-x^{-\alpha})$ for $x > 0$, which implies $\gamma = 1/2$ and $\rho = -1$,
- the standard Cauchy distribution leading to a tail behaviour with $\gamma = 1$ and $\rho = -2$,
- the Loggamma distribution with $\gamma = 1$ and $\rho = 0$ and density function

$$f(x) = \log(x) x^{-2} \, \mathbb{1}_{[1,\infty)}(x),$$

- the Burr distribution with a parametrisation such that $\gamma = 2$, $\rho = -1$ and distribution function

$$F(x) = 1 - (1 + \sqrt{x})^{-1}, \text{ for } x > 0,$$

- a logarithmically perturbed Pareto distribution of the random variable $g(U)$ with $\gamma = 1$ and $\rho = -1$, where $U \sim \text{Unif}(0, 1)$ and $g(x) = x^{-1}/\log(x^{-1})$. This distribution is denoted as negBias due to its negative bias in the Hill estimator.

On these distributions we evaluate the methods by their root mean square error (RMSE) when adaptively estimating γ with the Hill estimator relative to the RMSE obtained using k_{opt},

$$\text{EFF}_\gamma(\hat{k}) := \sqrt{\frac{\mathbb{E}_n[(\hat{\gamma}_{\hat{k}} - \gamma)^2]}{\mathbb{E}_n[(\hat{\gamma}_{k_{\text{opt}}} - \gamma)^2]}},$$

where $\mathbb{E}_n$ denotes the empirical expectation. These efficiency quotients are also used by, e.g., Guillou and Hall (2001), Gomes et al. (2000) and Drees and Kaufmann (1998). The smaller the quotient the better the threshold selection procedure performs compared to the asymptotically optimal sample fraction k_{opt}. Furthermore, we study the efficiency in quantile estimation with the estimator defined in (10) for $p = 0.001$,

$$\text{EFF}_q(\hat{k}) := \sqrt{\frac{\mathbb{E}_n[(\hat{q}_{\hat{k}} - q)^2]}{\mathbb{E}_n[(\hat{q}_{k_{\text{opt}}} - q)^2]}}.$$

Since we do not know the true minimizer k_{opt} of the AMSE, we utilize an empirical version suggested by Gomes et al. (2000). Following their approach we approximate k_{opt} by the mean of 20 independent replicates of $\bar{k}_{\text{opt}}$, which is the minimizer of the empirical MSE based on 1000 samples, i.e. $\bar{k}_{\text{opt}} = \underset{k}{\text{argmin}}\ \mathbb{E}_{n=1000}[(\hat{\gamma}_k - \gamma)^2]$.

We compare these efficiency values for eight different threshold selection methods. Most of the considered approaches are constructed for adaptive estimation of γ applying the Hill estimator. This includes one procedure that looks for a stable region among the Hill estimates, while the others aim to estimate k_{opt}. The only exception is the IHS approach discussed in Section 2, which is motivated to minimize the deviation from the exponential approximation. We still evaluate the performance of this procedure in the same simulations, although it is not primarily tailored for the specific applications. In total, the following methods are considered:

sIHS: IHS smoothed by using the *eBsc* package, see Section 2,

SAM: SAMSEE procedure with $\rho = -1$ as defined by (19) in Section 3,

GH: method by Guillou and Hall (2001) utilizing $c_{\text{crit}} = 1.25$ and $p = 1$,

DK: procedure by Drees and Kaufmann (1998) with fixed $\rho = -1$,

GO: approach by Goegebeur et al. (2008) defined in their equation (3.3) with fixed $\rho = -1$,

DB: double bootstrap approach by Danielsson et al. (2001) with the choice $n_1 = 120$ if $n = 500$ and $n_1 = 1000$ if $n = 5000$,

B: method by Beirlant et al. (2002) with $\rho = -1$,

RT: method by Reiss and Thomas (2007) with $\beta = 0$ as suggested by Neves and Fraga Alves (2004).

$n = 500$	SAM	GH	DK	GO	DB	B	RT	sIHS
Student-t(6)	1.07	1.68	1.18	1.38	1.06	1.04	1.04	1.14
Fréchet(2)	1.13	1.15	1.08	1.12	1.60	1.49	2.00	1.41
Cauchy	1.37	1.19	1.32	1.16	2.14	1.85	2.11	1.47
Loggamma	0.98	1.06	1.27	1.11	1.12	1.04	1.32	0.78
Burr(2,1)	1.11	1.22	1.47	1.13	1.68	1.42	1.82	1.14
negBias	1.06	1.13	1.56	1.13	1.07	1.22	1.89	2.27

$n = 5000$	SAM	GH	DK	GO	DB	B	RT	sIHS
Student-t(6)	1.20	1.58	1.31	1.39	1.35	1.03	1.26	1.03
Fréchet(2)	1.08	1.26	1.07	1.21	1.66	1.29	2.40	2.43
Cauchy	1.34	1.41	1.08	1.17	2.00	1.68	2.78	3.03
Loggamma	1.08	1.10	1.32	1.17	1.19	1.05	1.40	0.79
Burr(2,1)	1.07	1.29	1.62	1.14	1.63	1.29	2.21	1.79
negBias	0.98	1.12	1.54	1.10	1.30	1.04	2.08	3.98

Table 2: Efficiency values EFF_γ based on 2000 samples if $n = 500$ and on 500 samples if $n = 5000$. Lowest (best) efficiency values are highlighted in blue.

When looking at the results for estimating γ adaptively for $n = 500$ and $n = 5000$ in Table 2, we observe a very diverse picture of methods performing best. Overall we get the impression that SAMSEE together with the approach by Goegebeur et al. (2008) performs most stable over the variety of distributions. This is interesting, because those are the methods which depend least on tuning parameters. The performance of the approaches GH, DK and B is comparable, but we obtain from Table 2 that on average over all distributions the SAMSEE procedure is superior.
For estimating a high quantile SAMSEE also performs convincingly, see Table 3, but additionally sIHS and the approach by Danielsson et al. (2001) show very good efficiency values. They are closely followed by B and GO. Looking at the average performance over all distributions, SAMSEE performs best again. However, if we exclude the negBias distribution, sIHS

$n = 500$	SAM	GH	DK	GO	DB	B	RT	sIHS
Student-t(6)	1.09	2.30	1.23	1.60	1.01	1.04	1.16	1.10
Fréchet(2)	0.96	1.07	1.01	1.06	1.07	1.16	1.42	0.83
Cauchy	0.89	1.04	1.03	0.95	0.86	1.40	1.59	0.65
Loggamma	0.84	0.95	2.10	1.02	0.88	1.06	1.55	0.50
Burr(2,1)	0.79	2.15	8.60	0.98	0.71	1.43	3.19	0.41
negBias	1.66	1.37	2.32	2.13	0.80	1.98	3.75	8.05

$n = 5000$	SAM	GH	DK	GO	DB	B	RT	sIHS
Student-t(6)	1.07	1.39	1.16	1.29	1.44	1.02	1.24	0.94
Fréchet(2)	1.04	1.14	1.07	1.15	1.14	1.14	1.53	1.14
Cauchy	1.01	1.13	1.03	1.04	1.16	1.22	1.55	1.10
Loggamma	0.94	1.00	1.39	1.12	1.11	0.97	1.55	0.67
Burr(2,1)	1.00	1.38	1.24	1.11	1.09	1.11	1.72	0.71
negBias	1.00	1.11	0.87	1.21	0.87	1.21	1.23	2.16

Table 3: Efficiency values EFF_q for $p = 0.001$ based on 2000 samples if $n = 500$ and on 500 samples if $n = 5000$. Lowest (best) efficiency values are highlighted in blue.

works superior on average.

In conclusion, we can see that SAMSEE performs very efficiently and comparable to k_{opt} over all exemplary distributions. It works especially well for estimating a high quantile. Only in the case of estimating γ for the Cauchy distribution it performs worse than DK and GO, but still better than most other approaches. Recalling the results of the simulation on the influence of ρ in Table 1 in Section 3.1, it is not very surprising that SAMSEE performs slightly weaker in this situation. There, the Cauchy distribution is the only example we considered that benefits from estimating ρ instead of fixing it to -1.

From Table 3 we furthermore observe that sIHS is a strong choice when estimating high quantiles from small samples with n up to 5000. However, the performance when estimating γ is quite variable and it seems that sIHS does not perform particularly well for distributions with a small second order parameter ($\rho \leq -1$). This behaviour is already discussed in Section 2.1 and highlighted in Figure 2: sIHS selects smaller k than optimal for the Hill estimator, especially if ρ is in the regime between -1 and -8.

The reason why some approaches perform worse on quantiles than they do on γ is that the estimator $\hat{q}_k(p)$ defined in (10) depends on $\hat{\gamma}_k$ in the exponent and is thus very sensitive to overestimation in case of $\gamma > 1$. Hence,

when estimating a high quantile, an estimate $\hat{\gamma}_k$ that is too large will lead to an even stronger overestimation of the quantile. This is why a few outliers among the γ estimates can already cause much higher EFF_q values.

5 Application to varying extreme value index

In this section we analyse our new procedures in a financial application, where we study operational losses of a bank. We are, of course, particularly interested in the distributional properties of very high losses. It has been discussed before that it is reasonable to assume the distribution of such extreme losses being heavy-tailed (Chavez-Demoulin et al., 2016; Moscadelli, 2004) and to change with the financial market over time (Hambuckers et al., 2018; Cope et al., 2012). In this context, we want to estimate how the extreme value index changes depending on the univariate covariate time. For this task, we utilize the approaches presented in Sections 2 and 3 for locally optimal selection of a threshold.
The observations of interest are operational losses from the Italian bank UniCredit from 2005 to 2014. In Hambuckers et al. (2018) the data is analysed in a regularized generalized Pareto regression approach including several firm-specific, macroeconomic and financial indicators as covariates. This approach describes the dependence of the GPD parameters on various covariates via parametric functions.
We consider an easier and more direct approach to study the temporal dependence of the extreme value index without taking into account possible interference by other covariates. Our aim is to estimate the time dependent extreme value index $\gamma(t)$ non-parametrically with a simple ad hoc estimator that extends the estimator from de Haan and Zhou (2017) by employing our threshold selection procedures sIHS and SAMSEE. We present the estimator in Section 5.1 and the results we obtain when applying this estimator to the dataset of operational losses in Section 5.2.

5.1 Estimating a varying extreme value index

In de Haan and Zhou (2017), the authors already discussed estimating a trend in the extreme value index non-parametrically. They consider n independent random variables $X_i \sim F_{i/n}$, where $F_s \in \text{DoA}(G_{\gamma(s)})$ for $s \in [0, 1]$. To address this problem, they introduce the following estimator for $\gamma(s)$, which locally applies the Hill estimator and is based on a global sample

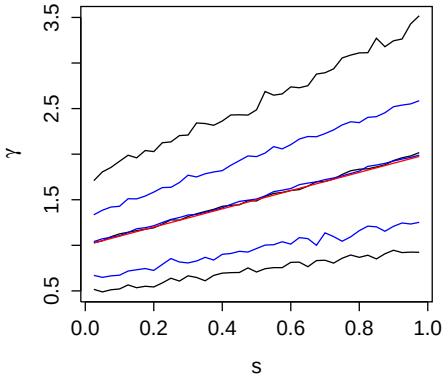


Figure 6: The true extreme value index $\gamma(s) = 1 + s$ (red) next to the averaged estimators over 1000 samples. The estimator from de Haan and Zhou (2017) is in black and its empirical 95% confidence interval is dashed. Our modified estimator employing SAMSEE is blue with dotted empirical confidence bounds.

fraction k,

$$\hat{\gamma}_k(s) := \frac{1}{2kh} \sum_{i \in I_n(s)} \left(\log X_i - \log X_{(\lfloor 2nh \rfloor - \lfloor 2kh \rfloor, \lfloor 2nh \rfloor)}\right)^+, \tag{22}$$

where $I_n(s)$ is the h-neighbourhood of s, i.e. $I_n(s) := \{i : |i/n - s| \leq h\}$. This estimator depends on the choice of the bandwidth h and the global sample fraction k, which is then rescaled to $2kh$ for the individual regions $I_n(s)$. A small bandwidth h leads to very high variability in $\hat{\gamma}_k(s)$ and a large value of h smooths out all interesting features. Thus, the choice of h should balance these two effects.
We suggest a modification of their estimator, where we locally estimate an optimal threshold $\hat{k}(s)$, i.e.

$$\hat{\gamma}_{\hat{k}(s)}(s) := \frac{1}{\hat{k}(s)} \sum_{i \in I_n(s)} \left(\log X_i - \log X_{(\lfloor 2nh \rfloor - \hat{k}(s), \lfloor 2nh \rfloor)}\right)^+. \tag{23}$$

To compare these two approaches, we repeat the simulation presented in Figure 2 (i) in de Haan and Zhou (2017) on samples of size $n = 5000$ with $X_i \sim$ Fréchet$(1/\gamma(i/n))$ and $\gamma(s) = 1+s$. Figure 6 illustrates the benefits of locally optimizing the threshold via SAMSEE from Section 3, as it strongly tightens the empirical confidence interval around the average, which is obtained from the 2.5% and 97.5% quantiles among 1000 estimates.

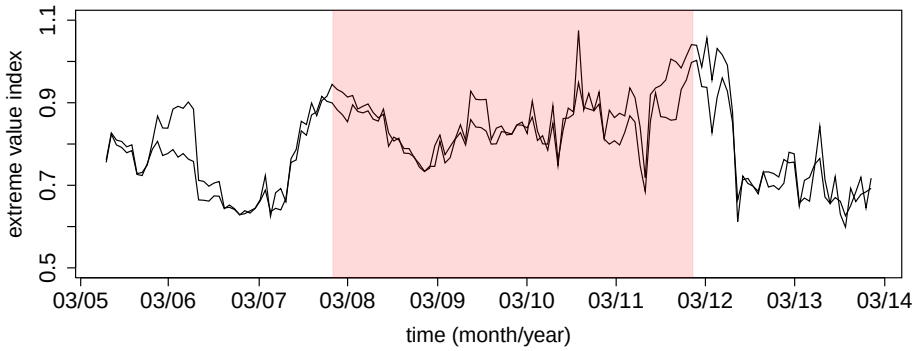


Figure 7: The non-parametric estimate of the extreme value index of the operational losses of type CPBP is presented using $\hat{k}$ from SAMSEE (black) and sIHS (dashed) and bandwidth $h = 0.05$. The red area indicates the time of the financial and Euro crisis.

5.2 Functional extreme value index of operational losses

The operational losses in the dataset of UniCredit are grouped by the type of event that caused the specific loss. We consider the event type CPBP, which provides sufficient observations for our local estimation approach. The CPBP losses are caused by clients, products and business practices related to derivatives or other financial instruments.
First we want to test if the extreme value index is constant over time. Using the test T4 from Einmahl et al. (2016), we can reject the null hypotheses with a p-value that is virtually zero and thus are confident that the extreme value index of the losses is indeed varying over time.
We apply the new methodology from (23) to these losses via estimating $\hat{k}$ with sIHS from Section 2 and the SAMSEE approach from Section 3. Figure 7 shows the estimates we obtain for the event type CPBP. It is clearly visible that both procedures yield similar estimates for most time points and that the simple ad hoc estimators recover an increase of the severity of high losses during the financial and Euro crisis from 2008 to 2011. A similar overall trend in the extreme value index can also be identified in the estimates of Hambuckers et al. (2018) for CPBP.
For a more extensive discussion of the data and results of the more complex model including further covariates we refer to Hambuckers et al. (2018).

A Theoretical results and proof of Theorem 3

Lemma 1 (Distribution of the Hill estimator)**.** *Let* $X_1, \dots, X_n \overset{i.i.d.}{\sim} F$ *for* $F \in \mathrm{DoA}(G_\gamma)$ *with* $\gamma > 0$*. Then the following distributional representation for the Hill estimator holds,*

$$\hat{\gamma}_k = \frac{1}{k} \sum_{i=1}^{k} \log \left(\frac{X_{(n-i+1,n)}}{X_{(n-k,n)}} \right) \overset{\mathcal{D}}{=} G_k + b_{n,k},$$

where $G_k \sim \Gamma(k, \gamma/k)$ *and for* $b_{n,k}$ *it holds that*

$$b_{n,k} \longrightarrow \begin{cases} 0, & \text{if } k/n \to 0, \\ b_c, & \text{if } k/n \to c, \end{cases} \qquad \text{as } n \to \infty,$$

for some $b_c \in \mathbb{R}$*.*

Proof. From the first order condition

$$\lim_{t \to \infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1/\gamma}$$

follows that $U = F^{\leftarrow}(1 - 1/x)$ is regularly varying with index γ and there exists a slowly varying function ℓ_U, such that $U(x) = x^\gamma \ell_U(x)$. Let $P_1, P_2, \dots$ be i.i.d. random variables with distribution function $1 - 1/y$. Note that $U(P_i) \overset{\mathcal{D}}{=} X_i$. We define $Y_{(k-i,k)} := \log(X_{(n-i,n)}) - \log(X_{(n-k,n)})$, for which it follows that

$$Y_{(k-i,k)} \overset{\mathcal{D}}{=} \log \left(\frac{U(P_{(n-i,n)})}{U(P_{(n-k,n)})} \right) = \gamma \log \left(\frac{P_{(n-i,n)}}{P_{(n-k,n)}} \right) + \log \left(\frac{\ell_U(P_{(n-i,n)})}{\ell_U(P_{(n-k,n)})} \right).$$

Note that $\log(P_i)$ is standard exponentially distributed. By Lemma 3.2.3 in de Haan and Ferreira (2006) follows for i.i.d. standard exponential random variables $E_1, E_2, \dots$ that $\left\{ E_{(n-i,n)} - E_{(n-k,n)} \right\}_{i=1}^{k-1} \overset{\mathcal{D}}{=} \left\{ E_{(k-i,k)} \right\}_{i=0}^{k-1}$.
Hence, we obtain for the Hill estimator that

$$\hat{\gamma}_k \overset{\mathcal{D}}{=} \gamma \frac{1}{k} \sum_{i=0}^{k-1} E_{(k-i,k)} + \frac{1}{k} \sum_{i=0}^{k-1} \log \left(\frac{\ell_U(P_{(n-i,n)})}{\ell_U(P_{(n-k,n)})} \right) \overset{\mathcal{D}}{=} G_k + b_{n,k},$$

where $G_k \sim \Gamma(k, \gamma/k)$ as the sum of i.i.d. exponentials and $b_{n,k}$ denotes the second average.
If $k/n \to 0$, $P_{(n-k,n)} \to \infty$ almost surely by Lemma 3.2.1 in de Haan and

Ferreira (2006). Since ℓ_U is slowly varying, $b_{n,k}$ converges to zero almost surely.
If $k/n \to c \in (0,1]$, $P_{(n-k,n)} \to 1/c$ in probability by Cor. 2.2.2 in de Haan and Ferreira (2006). Thus, by the weak law of large numbers

$$b_{n,k} \xrightarrow{\mathbb{P}} \mathbb{E}\left[\log\left(\frac{\ell_U(P)}{\ell_U(1/c)}\right)\Big|P > 1/c\right] =: b_c, \text{ as } n \to \infty.$$

□

Lemma 2. *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$ for $F \in \mathrm{DoA}(G_\gamma)$ with $\gamma > 0$. Then the following holds for $\mathrm{SKK} = (4-k)/(2\hat{\gamma}_k k)$ and $\mathrm{SKK}_- = (4+k)/(2\hat{\gamma}_k k)$ depending on the sample fraction k.*

1. *If k is finite,*
   - *then $\mathbb{E}[\mathrm{SKK}] + (2\gamma)^{-1} \to \frac{3}{2\gamma(k-1)} > 0$, as $n \to \infty$.*
   - *then $\mathbb{E}[\mathrm{SKK}^-] - (2\gamma)^{-1} \to \frac{5}{2\gamma(k-1)} > 0$, as $n \to \infty$.*
2. *If $k \to \infty,\ k/n \to 0$,*
   - *then $\mathbb{E}[\mathrm{SKK}] + (2\gamma)^{-1} \to 0$, as $n \to \infty$.*
   - *then $\mathbb{E}[\mathrm{SKK}^-] - (2\gamma)^{-1} \to 0$, as $n \to \infty$.*
3. *If $k \to \infty,\ k/n \to c > 0$*
   - *and $b_c \geq 0$, then $\mathbb{E}[\mathrm{SKK}] + (2\gamma)^{-1} \to \frac{b_c}{2\gamma(\gamma+b_c)} > 0$, as $n \to \infty$.*
   - *and $b_c \leq 0$, then $\mathbb{E}[\mathrm{SKK}^-] - (2\gamma)^{-1} \to \frac{-b_c}{2\gamma(\gamma+b_c)} > 0$, as $n \to \infty$.*

Proof. We proof these three alternative statements to obtain that the minimizing sequence $k = k_n$ is an intermediate sequence.

1. Let k be finite, by Lemma 1 holds that $\hat{\gamma}_k \overset{\mathcal{D}}{=} G_k + b_{k,n}$, where $G_k \sim \Gamma(k, \gamma/k)$ and $b_{k,n} \to 0$ as $n \to \infty$. Thus, as $n \to \infty$ it follows that

$$\mathbb{E}[\mathrm{SKK}] + (2\gamma)^{-1} \to \mathbb{E}\left[\frac{4-k}{2G_k k} + \frac{1}{2\gamma}\right] = \frac{3}{2\gamma(k-1)},$$
$$\mathbb{E}[\mathrm{SKK}^-] - (2\gamma)^{-1} \to \mathbb{E}\left[\frac{4+k}{2G_k k} - \frac{1}{2\gamma}\right] = \frac{5}{2\gamma(k-1)}.$$

2. The statement follows from the consistency of the Hill estimator and the continuous mapping theorem.

3. Let $k \to \infty$ and $k/n \to c > 0$, then it holds by Lemma 1 that $\hat{\gamma}_k \stackrel{\mathcal{D}}{=} G_k + b_{k,n}$, where $G_k \stackrel{\mathbb{P}}{\to} \gamma$ and $b_{k,n} \to b_c$ as $n \to \infty$. Thus, as $n \to \infty$

$$\mathbb{E}[\mathrm{SKK}] + (2\gamma)^{-1} \to \frac{-1}{2(\gamma + b_c)} + \frac{1}{2\gamma} = \frac{b_c}{2\gamma(\gamma + b_c)}, \text{ if } b_c > 0,$$
$$\mathbb{E}[\mathrm{SKK}^-] - (2\gamma)^{-1} \to \frac{1}{2(\gamma + b_c)} - \frac{1}{2\gamma} = \frac{-b_c}{2\gamma(\gamma + b_c)}, \text{ if } b_c < 0.$$

□

Lemma 3. *Let $E_1, \ldots, E_n$ be i.i.d. standard exponential random variables and k s.t. $1 \leq k \leq n$ and $k \to \infty$ as $n \to \infty$. We define the following random variables,*

$$P_{k,n} := \sqrt{k}\left(\frac{1}{k}\sum_{i=1}^{k} E_i - 1\right),$$
$$Q_{k,n} := \sqrt{k}\left(\frac{1}{k}\sum_{i=1}^{k} E_i^2 - 2\right),$$
$$R_{k,n} := \sqrt{k}\left(\frac{1}{k}\sum_{i=1}^{k} e_{i+1}^k E_i - 1\right),$$

where $e_i^k := \sum_{l=i}^{k} l^{-1} = \mathbb{E}[E_{(k-i+1,k)}]$. Then it holds for $n \to \infty$ that

$$(P_{k,n}, Q_{k,n}, R_{k,n})^T \xrightarrow{\mathcal{D}} \mathcal{N}\left((0,0,0)^T, \begin{pmatrix} 1 & 4 & 1 \\ 4 & 20 & 4 \\ 1 & 4 & 2 \end{pmatrix}\right).$$

Proof. We use the Cramér-Wold device that gives us a joint normal limit distribution if all linear combinations have an univariate normal limit distribution. For $a_1, a_2, a_3 \in \mathbb{R}$ we study

$$(a_1 P_{k,n} + a_2 Q_{k,n} + a_3 R_{k,n}). \tag{24}$$

To prove asymptotic normality for the sum in (24) we use Liapounov's central limit theorem (CLT), see Theorem 7.1.2. in Chung (1974). We consider a sum $S_n := \sum_{i=1}^{k} X_{i,k}$ of independent random variables fulfilling the following three conditions,

1) $\mathbb{E}[X_{i,k}] = 0,\ \forall k\ \forall i,$

2) $\sum_{i=1}^{k} \mathrm{Var}(X_{i,k}) = \sigma^2,$

3) $\Gamma(k) = \sum_{i=1}^{k} \mathbb{E}[|X_{i,k}|^3] \longrightarrow 0,$ as $k \to \infty,$

Then the CLT proves a standard normal limit for S_n. We define

$$X_{i,k} := \frac{1}{\sqrt{k}} \Big(a_1 E_i - a_1 + a_2 E_i^2 - 2a_2 + a_3 e_{i+1}^k E_i - a_3 e_{i+1}^k \Big)$$

for $i = 1, \ldots, k$ where $e_{k+1}^k := 0$, such that $\sum_{i=1}^{k} X_{i,k} \approx (24)$, where $c_k \approx c$ if $c_k \to c$ as $k \to \infty$ and the approximation error is due to

$$\frac{1}{k} \sum_{i=1}^{k-1} e_{i+1}^k = \frac{1}{k} \sum_{i=1}^{k-1} \frac{1}{k} \sum_{l=i+1}^{k} \frac{1}{l/k} \ \approx\ \int_{\frac{1}{k}}^{1} \int_{v}^{1} \frac{1}{u} \mathrm{d}u \mathrm{d}v = -\int_{\frac{1}{k}}^{1} \log(v) \mathrm{d}v \underset{k\to\infty}{\to} 1.$$

Now we have to check the three conditions. Condition 1) follows immediately from $\mathbb{E}[E_i] = 1$ and $\mathbb{E}[E_i^2] = 2$. For condition 2) we need to calculate the variance

$$\begin{aligned}
\mathrm{Var}(X_{i,k}) &= \mathrm{Var}\left(\frac{1}{\sqrt{k}} \Big(a_1 E_i - a_1 + a_2 E_i^2 - 2a_2 + a_3 e_i^k E_i - a_3 e_i^k \Big) \right) \\
&= \frac{1}{k} \Big(a_1^2 \mathrm{Var}(E_i) + a_2^2 \mathrm{Var}(E_i^2) + a_3^2 (e_{i+1}^k)^2 \mathrm{Var}(E_i) \\
&\qquad + 2(a_1 a_2 + a_2 a_3 e_{i+1}^k) \,\mathrm{Cov}(E_i, E_i^2) + 2a_1 a_3 \mathrm{Var}(E_i) \Big) \\
&= \frac{1}{k} \Big(a_1^2 + 20 a_2^2 + a_3^2 (e_{i+1}^k)^2 + 8(a_1 a_2 + a_2 a_3 e_{i+1}^k) + 2a_1 a_3 \Big),
\end{aligned}$$

since $\mathrm{Var}(E_i) = 1$, $\mathrm{Var}(E_i^2) = 20$ and $\mathrm{Cov}(E_i, E_i^2) = \mathbb{E}[E_i^3] - \mathbb{E}[E_i^2]\mathbb{E}[E_i] = 4$. With the approximation

$$\frac{1}{k} \sum_{i=1}^{k-1} (e_{i+1}^k)^2 \ \approx\ \int_{\frac{1}{k}}^{1} \Big(\int_{v}^{1} \frac{1}{u} \mathrm{d}u \Big)^2 \mathrm{d}v \underset{k\to\infty}{\to} 2$$

follows that

$$\sum_{i=1}^{k} \mathrm{Var}(X_{i,k}) = a_1^2 + 20 a_2^2 + 2a_3^2 + 8a_1 a_2 + 8a_2 a_3 + 2a_1 a_3.$$

Condition 3) holds with

$$\Gamma(k) = \frac{1}{k\sqrt{k}} \sum_{i=1}^{k} \mathbb{E}\Big[\big| (a_1 + a_3 e_i^k)(E_i - 1) + a_2 (E_i^2 - 2) \big|^3 \Big] = \frac{c}{\sqrt{k}} \to 0, \quad (25)$$

as $k \to \infty$ and for a constant $c > 0$, since the exponential distribution has finite moments and $\sum_{i=1}^{k} (e_{i+1}^k)^3/k \approx 6$. Thus, we obtain that

$$\sum_{i=1}^{k} X_{i,k} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\ a_1^2 + 20a_2^2 + 2a_3^2 + 8a_1a_2 + 8a_2a_3 + 2a_1a_3).$$

This is the limiting distribution of the sum in (24) and also follows from the joint normal distribution. □

Lemma 4. *Let* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$ *for* $F \in \mathrm{DoA}(G_\gamma)$ *with* $\gamma > 0$ *and* $P_1, P_2, \ldots$ *be i.i.d. random variables with distribution function* $1 - 1/y$. *We define*

$$\hat{\gamma}_k := \frac{1}{k} \sum_{i=1}^{k} \log \left(\frac{X_{(n-i+1,n)}}{X_{(n-k,n)}} \right), \quad M_n := \frac{1}{k} \sum_{i=1}^{k} \log \left(\frac{X_{(n-i+1,n)}}{X_{(n-k,n)}} \right)^2,$$
$$\overline{YE} := \frac{1}{k} \sum_{i=1}^{k} \log \left(\frac{X_{(n-i+1,n)}}{X_{(n-k,n)}} \right) e_i^k, \ \textit{and} \quad e_i^k := \sum_{l=i}^{k} \frac{1}{l}.$$

If the second order condition

$$\lim_{t \to \infty} \frac{\frac{U(tx)}{U(t)} - x^\gamma}{A(t)} = x^\gamma \frac{x^\rho - 1}{\rho}$$

holds for $\rho < 0$ *and* $x > 0$, *then*

$$\hat{\gamma}_k \overset{\mathcal{D}}{=} \gamma + \gamma P_{k,n}/\sqrt{k} + \frac{A(Y_{(n-k,n)})}{1-\rho} + o_p(A(n/k)), \tag{26}$$

$$M_n \overset{\mathcal{D}}{=} 2\gamma^2 + \gamma^2 Q_{k,n}/\sqrt{k} + \frac{2\gamma(2-\rho)}{(1-\rho)^2} A(Y_{(n-k,n)}) + o_p(A(n/k)), \tag{27}$$

$$\overline{YE} \overset{\mathcal{D}}{=} 2\gamma + \gamma(P_{k,n} + R_{k,n})/\sqrt{k} + \frac{2-\rho}{(1-\rho)^2} A(Y_{(n-k,n)}) + o_p(A(n/k)). \tag{28}$$

Proof. The results in (26) and (27) are already stated in the proof of Theorem 1 in de Haan and Peng (1998).
To prove (28) we follow the proof of the asymptotic normality of the Hill estimator in de Haan and Ferreira (2006). Let A_0 be such that $A(t)/A_0(t) \to 1$, as $t \to \infty$. Then, for each $\epsilon > 0$ there exists a t_0 such that for $t \geq t_0$ and $x \geq 1$ the inequality in Theorem B.2.18 in de Haan and Ferreira (2006)

holds. For $t = P_{n-k,n}$ and $x = P_{(n-i,n)}/P_{(n-k,n)}$ we obtain that

$$\overline{YE} \stackrel{\mathcal{D}}{=} \frac{\gamma}{k}\sum_{i=1}^{k}\log\left(\frac{P_{(n-i+1,n)}}{P_{(n-k,n)}}\right)e_i^k + A_0(P_{(n-k,n)})\frac{1}{k}\sum_{i=1}^{k}\frac{\left(\frac{P_{(n-i+1,n)}}{P_{(n-k,n)}}\right)^{\rho}-1}{\rho}e_i^k$$
$$+ o_p(1)|A_0(P_{(n-k,n)})|\frac{1}{k}\sum_{i=1}^{k}\left(\frac{P_{(n-i+1,n)}}{P_{(n-k,n)}}\right)^{\rho+\epsilon}e_i^k.$$

The second term can be approximated by

$$\frac{1}{k}\sum_{i=1}^{k}\frac{\left(\frac{P_{(n-i+1,n)}}{P_{(n-k,n)}}\right)^{\rho}-1}{\rho}e_i^k \to \int_0^1 \frac{v^{-\rho}-1}{\rho}\int_v^1 \frac{1}{u}\mathrm{d}u\mathrm{d}v = \frac{2-\rho}{(1-\rho)^2},$$

and for the third term holds

$$\frac{1}{k}\sum_{i=1}^{k}\left(\frac{P_{(n-i+1,n)}}{P_{(n-k,n)}}\right)^{\rho+\epsilon}e_i^k \to \int_0^1 v^{-\rho-\epsilon}\int_v^1 \frac{1}{u}\mathrm{d}u\mathrm{d}v = \frac{1}{(1-\rho-\epsilon)^2},$$

as $k \to \infty$. Note that for $E_1, \ldots, E_n$ i.i.d. standard exponential random variables follows by Rényi's representation that

$$\left\{\log\left(\frac{P_{(n-i+1,n)}}{P_{(n-k,n)}}\right)\right\}_{i=1}^{k} \stackrel{\mathcal{D}}{=} \left\{E_{(k-i+1,k)}\right\}_{i=1}^{k} \stackrel{\mathcal{D}}{=} \left\{\sum_{j=i}^{k}\frac{E_j}{j}\right\}_{i=1}^{k}.$$

This distributional equality enables the following transformations,

$$\sum_{i=1}^{k}\log\left(\frac{P_{(n-i+1,n)}}{P_{(n-k,n)}}\right)e_i^k \stackrel{\mathcal{D}}{=} \sum_{i=1}^{k}e_i^k\sum_{j=i}^{k}\frac{E_j}{j}$$
$$= \sum_{i=1}^{k}\frac{E_i}{i}\sum_{j=1}^{i}e_j^k = \sum_{i=1}^{k}\frac{E_i}{i}\Big(ie_{i+1}^k + \sum_{j=1}^{i}\sum_{l=j}^{i}\frac{1}{l}\Big)$$
$$= \sum_{i=1}^{k}\frac{E_i}{i}\Big(i + ie_{i+1}^k\Big) = \sum_{i=1}^{k}E_i + \sum_{i=1}^{k}E_ie_{i+1}^k.$$

Thus,

$$\frac{\gamma}{k}\sum_{i=1}^{k}\log\left(\frac{P_{(n-i+1,n)}}{P_{(n-k,n)}}\right)e_i^k \stackrel{\mathcal{D}}{=} 2\gamma + \gamma\Big(\Big(\frac{1}{k}\sum_{i=1}^{k}E_i - 1\Big) + \Big(\frac{1}{k}\sum_{i=1}^{k}E_ie_{i+1}^k - 1\Big)\Big)$$
$$= 2\gamma + \gamma\left(P_{k,n} + R_{k,n}\right)/\sqrt{k}.$$

Combining the above arguments as in the proof of Theorem 3.2.5 in de Haan and Ferreira (2006) gives (28). □

Lemma 5. *For $k \to \infty$, $k/n \to 0$ and $k/K \to c$ with $0 < c < 1$,*

$$\begin{aligned}\mathrm{Cov}(R_{K,n}, R_{k,n}) &\to \frac{2c - c\log(c)}{\sqrt{c}},\\ \mathrm{Cov}(R_{K,n}, P_{k,n}) &\to \frac{c - c\log(c)}{\sqrt{c}}, \text{ as } n \to \infty,\end{aligned}$$

where $R_{k,n}$ and $P_{k,n}$ are defined in Lemma 3.

Proof. Let $E_1, E_2, \ldots$ be i.i.d. standard exponential random variables, where $\mathrm{Cov}(E_i, E_j)$ is equal to 1 if $i = j$ and 0 otherwise. Then

$$\begin{aligned}\mathrm{Cov}(R_{K,n}, R_{k,n}) &= \mathrm{Cov}\left(\sum_{i=1}^{k} E_i \frac{e_{i+1}^k}{\sqrt{k}}, \sum_{i=1}^{K} E_i \frac{e_{i+1}^K}{\sqrt{K}}\right)\\ &= \sum_{i=1}^{k}\sum_{j=1}^{K} \frac{e_{i+1}^k}{\sqrt{k}} \frac{e_{j+1}^K}{\sqrt{K}} \mathrm{Cov}(E_i, E_j) = \frac{\sqrt{k}}{\sqrt{K}} \frac{1}{k} \sum_{i=1}^{k} \left(\frac{1}{k}\sum_{l=i+1}^{k} \frac{1}{l/k}\right)\left(\frac{1}{K}\sum_{l=i+1}^{K} \frac{1}{l/K}\right)\\ &\approx \frac{\sqrt{k}}{\sqrt{K}} \frac{1}{k} \sum_{i=1}^{k} \left(\int_{(i+1)/k}^{1} \frac{1}{u}\mathrm{d}u\right)\left(\int_{(i+1)/K}^{1} \frac{1}{u}\mathrm{d}u\right)\\ &\approx \sqrt{c}\int_0^1 \log(v)^2 - \log(v)\log(c)\mathrm{d}v = 2\sqrt{c} - \sqrt{c}\log(c),\end{aligned}$$

where $c_k \approx c$ again denotes that $c_k \to c$ as $k \to \infty$.
In the same way we obtain

$$\begin{aligned}\mathrm{Cov}(R_{K,n}, P_{k,n}) &= \sum_{i=1}^{k}\sum_{j=1}^{K} \frac{1}{\sqrt{k}} \frac{e_{j+1}^K}{\sqrt{K}} \mathrm{Cov}(E_i, E_j)\\ &\approx \frac{\sqrt{k}}{\sqrt{K}} \frac{1}{k} \sum_{i=1}^{k} \left(\int_{(i+1)/K}^{1} \frac{1}{u}\mathrm{d}u\right) \approx \frac{\sqrt{k}}{\sqrt{K}} \int_0^1 \left(\int_{cv}^{1} \frac{1}{u}\mathrm{d}u\right)\mathrm{d}v\\ &= \sqrt{c} - \sqrt{c}\log(c)\end{aligned}$$

□

Theorem 4. *Let $\bar{\gamma}_k := \frac{1}{k}\sum_{i=1}^{k} \hat{\gamma}_i$ denote the average over Hill estimates. Further let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution function*

$F \in \mathrm{DoA}(G_\gamma)$, $\gamma > 0$. *If*

$$\lim_{t\to\infty} \frac{\frac{U(tx)}{U(t)} - x^\gamma}{A(t)} = x^\gamma \frac{x^\rho - 1}{\rho}, \tag{29}$$

with $U(x) := F^{\leftarrow}\left(1 - \frac{1}{x}\right)$ *holds and* $k \to \infty$ *and* $k/n \to 0$ *as* $n \to \infty$,

$$\sqrt{k}\left(\bar{\gamma}_k - \gamma\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\frac{\lambda}{(1-\rho)^2}, 2\gamma^2\right).$$

with $\lambda := \lim_{k\to\infty} \sqrt{k} A(n/k)$.

Proof. First we have to rewrite the average over the Hill estimator,

$$\begin{aligned}
\bar{\gamma}_k &= \frac{1}{k}\sum_{i=1}^{k} \hat{\gamma}_k = \frac{1}{k}\sum_{i=1}^{k} \frac{1}{i}\sum_{j=1}^{i} \log\left(\frac{X_{(n-j+1,n)}}{X_{(n-i,n)}}\right) \\
&= \frac{1}{k}\sum_{i=1}^{k} \log X_{(n-i+1,n)} \sum_{j=i}^{k} \frac{1}{j} - \frac{1}{k}\sum_{i=1}^{k} \log X_{(n-i,n)} \\
&= \frac{1}{k}\sum_{i=1}^{k} \log\left(\frac{X_{(n-i+1,n)}}{X_{(n-k,n)}}\right) \sum_{j=i}^{k} \frac{1}{j} - \frac{1}{k}\sum_{i=1}^{k} \log\left(\frac{X_{(n-i,n)}}{X_{(n-k,n)}}\right) \\
&= \overline{YE} - \hat{\gamma}_k + \frac{1}{k}\log\left(\frac{X_{(n,n)}}{X_{(n-k,n)}}\right),
\end{aligned}$$

where $\overline{YE}$ is defined in Lemma 4. Following the proof of Lemma 4 it holds that the last term above is in distribution equal to

$$\frac{\gamma}{k}\log\left(\frac{P_{(n,n)}}{P_{(n-k,n)}}\right) + \frac{A(P_{(n-k,n)})}{k}\frac{\left(\frac{P_{(n,n)}}{P_{(n-k,n)}}\right)^\rho - 1}{\rho} + o_p(1)\frac{|A(P_{(n-k,n)})|}{k}\left(\frac{P_{(n,n)}}{P_{(n-k,n)}}\right)^{\rho+\epsilon}.$$

From Corollary 2.2.2 in de Haan and Ferreira (2006) follows that

$$\frac{k}{n}P_{(n-k,n)} \xrightarrow{\mathbb{P}} 1, \text{ and } \frac{P_{(n,n)}}{kP_{(n-k,n)}} \xrightarrow{\mathbb{P}} 1, \text{ as } n \to \infty.$$

Thus, $\big(\log(X_{(n,n)}) - \log(X_{(n-k,n)})\big)/k = O_p(\log(k)/k) + O_p(A(n/k)/k)$, and by Lemma 4 follows

$$\bar{\gamma}_k \overset{\mathcal{D}}{=} \gamma + \gamma R_{k,n}/\sqrt{k} + \frac{A(n/k)}{(1-\rho)^2} + o_p(A(n/k)) + O_p(\log(k)/k).$$

□

Theorem 5. *Let* $\bar{\gamma}_{\text{up},K,k} := \frac{1}{K-k+1}\sum_{i=k}^{K}\hat{\gamma}_i$ *be the upper mean and* k *and* K *intermediate sequences, i.e.* $k \to \infty$ *and* $k/n \to 0$*, as* $n \to \infty$*. Further, let* $\sqrt{k}A(n/k) \to \lambda$ *and* $k/K \to c$ *with* $0 < c < 1$*. Under the conditions of Theorem 4, it holds that*

$$\sqrt{k}(\bar{\gamma}_{\text{up},K,k} - \gamma) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\frac{\lambda}{(1-\rho)^2}\frac{c^\rho - c}{1-c}, \frac{2\gamma^2 c}{1-c}\Big(1 + \frac{c\log(c)}{1-c}\Big)\right).$$

Proof. We can write the upper mean as a combination of two averaged Hill estimators and apply Theorem 4,

$$\begin{aligned}\bar{\gamma}_{\text{up},K,k} &= \frac{K}{K-k+1}\bar{\gamma}_K - \frac{k}{K-k+1}\bar{\gamma}_k \\ &\overset{\mathcal{D}}{=} \gamma + \gamma\frac{K}{K-k+1}R_{K,n}/\sqrt{K} - \gamma\frac{k}{K-k+1}R_{k,n}/\sqrt{k} \\ &\quad + \frac{K}{K-k+1}\frac{A(n/K)}{(1-\rho)^2} - \frac{k}{K-k+1}\frac{A(n/k)}{(1-\rho)^2} + o_p(A(n/K)).\end{aligned}$$

We approximate k/K by c and obtain

$$\begin{aligned}\sqrt{k}(\bar{\gamma}_{\text{up},K,k} - \gamma) &\overset{\mathcal{D}}{=} \gamma\frac{\sqrt{c}}{1-c}R_{K,n} - \gamma\frac{c}{(1-c)}R_{k,n} \\ &\quad + \frac{1}{1-c}\frac{\sqrt{k}A(n/K)}{(1-\rho)^2} - \frac{c}{(1-c)}\frac{\sqrt{k}A(n/k)}{(1-\rho)^2} + o_p(1).\end{aligned}$$

Now we need the covariance between $R_{n,k}$ and $R_{n,K}$, see Lemma 5, and apply the following property of regular varying functions,

$$\sqrt{k}A(n/K) = \sqrt{k}A(n/k)\frac{A(cn/k)}{A(n/k)} \to \lambda c^\rho, \text{ as } n \to \infty.$$

This leads to

$$\begin{aligned}&\sqrt{K}(\bar{\gamma}_{\text{up},K,k} - \gamma) \\ &\quad \xrightarrow{\mathcal{D}} \mathcal{N}\left(\frac{\lambda}{(1-\rho)^2}\frac{c^\rho - c}{1-c}, \frac{\gamma^2 c}{(1-c)^2}\Big(2 + 2c - 2\sqrt{c}\Big(\frac{2c - c\log(c)}{\sqrt{c}}\Big)\Big)\right).\end{aligned}$$

□

Proof of Theorem 3. The bias estimator is defined as $\bar{b}_{\text{up},K,k} = \bar{\gamma}_{\text{up},K,k} - \bar{\gamma}_K$ in equation (14). Thus, we can utilize the asymptotic normality results for

$\bar{\gamma}_k$ and $\bar{\gamma}_{\mathrm{up},K,k}$ in Theorem 4 and 5. Following the proofs of these theorems it holds that

$$\sqrt{k}\,\bar{b}_{\mathrm{up},K,k} \stackrel{\mathcal{D}}{=} \gamma\frac{k-1}{K-k+1}\frac{\sqrt{k}}{\sqrt{K}}R_{K,n} - \gamma\frac{k}{K-k+1}R_{k,n} + \frac{k}{K-k+1}\frac{\sqrt{k}(A(n/K)-A(n/k))}{(1-\rho)^2} + \sqrt{k}\Big(o_p(A(n/K)) + o_p(A(n/k))\Big).$$

Here the random variable $R_{k,n}$ is defined in Lemma 3 and we know that $R_{k,n}$ has a normal limit distribution. With Lemma 3 and Lemma 5 we obtain the following variance,

$$\begin{aligned}
&\mathrm{Var}\left(\gamma\frac{k-1}{K-k+1}\frac{\sqrt{k}}{\sqrt{K}}R_{K,n} - \gamma\frac{k}{K-k+1}R_{k,n}\right)\\
&\approx \gamma^2\left(\left(\frac{c\sqrt{c}}{1-c}\right)^2\mathrm{Var}(R_{K,n}) + \left(\frac{c}{1-c}\right)^2\mathrm{Var}(R_{k,n}) - 2\frac{c^2\sqrt{c}}{(1-c)^2}\,\mathrm{Cov}(R_{n,K},R_{n,k})\right)\\
&\approx \gamma^2\frac{2c^3+2c^2-4c^3+2c^3\log(c)}{(1-c)^2} = \frac{2\gamma^2c^2}{1-c}\left(1+\frac{c\log(c)}{1-c}\right),
\end{aligned}$$

The bias term of the normal limit is

$$\frac{k}{K-k+1}\frac{\sqrt{k}(A(n/K)-A(n/k))}{(1-\rho)^2} \to \frac{\lambda}{(1-\rho)^2}\frac{c(c^\rho-1)}{1-c},$$

as $n\to\infty$, which follows from the regular variation of A and due to $\sqrt{k}A(n/k)\to\lambda$. Since

$$\sqrt{k}\Big(o_p(A(n/K)) + o_p(A(n/k))\Big) = o_p(1),$$

the statement of the theorem follows immediately. □

Acknowledgement

Support of the DFG RTG 2088 (B4) is gratefully acknowledged. We are grateful to Julien Hambuckers for providing us the dataset of operational losses from UniCredit.

References

Bader, B., J. Yan, and X. Zhang (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *Ann. Appl. Stat. 12*, 310–329.

Beirlant, J., G. Dierckx, A. Guillou, and C. Stărică (2002). On exponential representations of log-spacings of extreme order statistics. *Extremes 5*, 157–180.

Beirlant, J., A. Kijko, T. Reynkens, and J. H. J. Einmahl (2018). Estimating the maximum possible earthquake magnitude using extreme value methodology: the Groningen case. *Nat. Hazards 169*, 1–23.

Carreau, J., P. Naveau, and L. Neppel (2017). Partitioning into hazard subregions for regional peaks-over-threshold modeling of heavy precipitation. *Water Resour. Res. 53*, 4407–4426.

Chavez-Demoulin, V., P. Embrechts, and M. Hofert (2016). An extreme value approach for modeling operational risk losses depending on covariates. *J. Risk Insur. 83*, 735–776.

Chung, K. L. (1974). *A Course in Probability Theory*. Academic Press.

Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009). Power-law distributions in empirical data. *SIAM Rev.*, 661–703.

Cope, E. W., M. T. Piche, and J. S. Walter (2012). Macroenvironmental determinants of operational loss serverity. *J. Bank. Finance 36*, 1362–1380.

Danielsson, J., L. de Haan, L. Peng, and C. G. de Vries (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *J. Multivariate Anal. 76*, 226–248.

de Haan, L. and A. Ferreira (2006). *Extreme Value Theory - An Introduction*. Springer.

de Haan, L. and L. Peng (1998). Comparison of tail index estimators. *Stat. Neerl. 52*, 60–70.

de Haan, L. and C. Zhou (2017). Trends in extreme value indices. working paper, https://personal.eur.nl/zhou/Research/WP/varygamma.pdf.

Dey, D. Y. and J. Yan (2016). *Extreme value modeling and risk analysis*. Taylor and Francis Group.

Drees, H., A. Janßen, S. I. Resnick, and T. Wang (2018). On a minimum distance procedure for threshold selection in tail analysis. preprint, arXiv:1811.06433.

Drees, H. and E. Kaufmann (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stoch. Proc. Appl. 75*, 149–172.

Drees, H., S. Resnick, and L. de Haan (2000). How to make a Hill plot. *Ann. Statist. 28*, 254–274.

DuMouchel, W. H. (1983). Estimating the stable index α in order to measure tail thickness: A critique. *Ann. Statist. 11*, 1019–1031.

Einmahl, J. H. J., L. de Haan, and C. Zhou (2016). Statistics of heteroscedastic extremes. *J. R. Statist. Soc B 78*, 31–51.

Ferreira, A., L. de Haan, and L. Peng (2003). On optimising the estimation of high quantiles of a probability distribution. *Statistics 37*, 401–434.

Goegebeur, Y., J. Beirlant, and T. de Wet (2008). Linking Pareto-tail kernel goodness-of-fit statistics with tail index at optimal threshold and second order estimation. *REVSTAT - Stat. J. 6*, 51–69.

Gomes, M. I., M. J. a. Martins, and M. Neves (2000). Alternatives to a semi-parametric estimator of parameters of rare events - the Jackknife methodology. *Extremes 3*, 207–229.

Gonzalo, J. and J. Olmo (2004). Which extreme values are really extreme? *J. Financ. Econom. 2*, 349–369.

Guillou, A. and P. Hall (2001). A diagnostic for selecting the threshold in extreme value analysis. *J. R. Statist. Soc. B 63*, 293–305.

Hall, P. and I. Weissman (1997). On the estimation of extreme tail probebilities. *Ann. Statist. 25*, 1311–1326.

Hambuckers, J., A. Groll, and T. Kneib (2018). Understanding the economic determinants of the serverity of operational losses: a regularized generalized Pareto regression approach. *J. Appl. Econom. 33*, 898–935.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist. 3*, 1163–1174.

Kratz, M. F. and S. I. Resnick (1996). The qq-estimator and heavy tails. *Comm. Statist. Stochastic Models 12*, 699–724.

Krivobokova, T. and G. Kauermann (2007). A note on penalized spline smoothing with correlated errors. *J. Am. Statist. Ass. 102*, 1328–1337.

Lee, Y. K., E. Mammen, and B. U. Park (2010). Bandwidth selection for kernel regression with correlated errors. *Statistics 44*, 327–340.

Moscadelli, M. (2004). The modelling of operational risk: experience with the analysis of the data collected by the basel committee. Technical report, Bank of Italy – Banking and Finance Department.

Neves, C. and M. I. Fraga Alves (2004). Reiss and Thomas' automatic selection of the number of extremes. *Comput. Stat. Data Anal. 47*, 689–704.

Opsomer, J., Y. Wang, and Y. Yang (2001). Nonparametric regression with correlated errors. *Statist. Sci. 16*, 134–153.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist. 3*, 119–131.

Reiss, R.-D. and M. Thomas (2007). *Statistical Analysis of Extreme Values*. Birkhäuser Verlag.

Resnick, S. and C. Stărică (1997). Smoothing the Hill estimator. *Adv. Appl. Probab. 29*, 271–293.

Scarrott, C. and A. MacDonald (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT - Stat. J. 10*, 33–60.

Serra, P., T. Krivobokova, and F. Rosales (2018). Adaptive non-parametric estimation of mean and autocovariance in regression with dependent errors. preprint, arXiv:1812.06948.