The Annals of Statistics
2021, Vol. 49, No. 6, 3534–3558
https://doi.org/10.1214/21-AOS2096

POSTERIOR ANALYSIS OF n IN THE BINOMIAL (n, p) PROBLEM WITH BOTH PARAMETERS UNKNOWN—WITH APPLICATIONS TO QUANTITATIVE NANOSCOPY

BY JOHANNES SCHMIDT-HIEBER[1], LAURA FEE SCHNEIDER[2,*], THOMAS STAUDT[2,4,†], ANDREA KRAJINA[2,‡], TIMO ASPELMEIER[2,§] AND AXEL MUNK[2,3,4,¶]

[1]*Department of Applied Mathematics, University of Twente, a.j.schmidt-hieber@utwente.nl*

[2]*Institute for Mathematical Stochastics, University of Göttingen , *laura-fee.schneider@mathematik.uni-goettingen.de; †thomas.staudt@uni-goettingen.de; ‡andrea.krajina@mathematik.uni-goettingen.de; §timo.aspelmeier@mathematik.uni-goettingen.de; ¶munk@math.uni-goettingen.de*

[3]*Max Planck Institute for Biophysical Chemistry*

[4]*Cluster of Excellence "Multiscale Bioimaging: from Molecular Machines to Networks of Excitable Cells" (MBExC), University of Göttingen, University Medical Center*

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. For a large class of priors, we establish posterior contraction and a Bernstein-von Mises type theorem in a setting where $p \to 0$ and $n \to \infty$ as $k \to \infty$. Furthermore, we suggest a new class of Bayesian estimators for n and provide a comprehensive simulation study in which we investigate their performance. To showcase the advantages of a Bayesian approach on real data, we also benchmark our estimators in a novel application from super-resolution microscopy.

1. Introduction. The binomial distribution with parameters n and p is the most fundamental model for the repetition of independent success/failure events. Motivated by several important applications, we focus on the situation where both p and n are unknown. For example, n might 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 (Betzig et al. (2006), Hell (2009), Aspelmeier, Egner and Munk (2015)).

The 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 at least to Fisher (1941). In comparison to the estimation of one of the parameters when the other is known (Lehmann and Casella (1996)), this problem turns out to be much harder. Fisher, who regarded the assumption of an unknown integer n as "entirely academic", suggested the use of the sample maximum, arguing that this estimator is necessarily good if the sample size is sufficiently large. Indeed, if $X_1, \dots, X_k$ are i.i.d. $\mathrm{Bin}(n, p)$ distributed random variables for fixed $n \in \mathbb{N}$ and $p \in (0, 1)$, the sample maximum $M_k := \max_{i=1,\dots,k} X_i$ converges exponentially fast to n as $k \to \infty$, since

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

Received January 2019; revised May 2021.
MSC2020 subject classifications. Primary 62G05; secondary 62F15, 62F12, 62P10, 62P35.
Key words and phrases. Bayesian estimation, posterior contraction, Bernstein-von Mises theorem, binomial distribution, beta-binomial likelihood, quantitative cell imaging.

In practice, however, the regime with small p ("rare events") is often the relevant one (see the references below and Section 4). In this setting, the sample maximum strongly underestimates the true n even for large sample sizes k. 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 $\mathbb{P}(M_k \geq n/2) \geq 1/2$. If $p = 0.1$ and $n = 20$, one would even need a sample size of more than $k = 900{,}000$ for the same probability.

The erratic behavior of the sample maximum can be explored by allowing the parameters n and p to depend on k. Applying Bernoulli's inequality and the bound $1 - x \leq e^{-x}$, it follows from (1.1) that $1 - e^{-kp^n} \leq \mathbb{P}(M_k = n) \leq kp^n$. Therefore, the sample maximum M_k becomes an inconsistent estimator of n if $kp^n \to 0$ as $k \to \infty$ (see Lemma A.1 in the Supplementary Material (Schmidt-Hieber et al. (2021)) for a characterization of domains of consistency and inconsistency of M_k). One particular example where consistency breaks down is the domain of attraction of the Poisson distribution: when $n \geq \log(k) \to \infty$ and $p \to 0$ such that $np \to \mu \in (0, \infty)$, then $kp^n \leq k^{1+\log p} \to 0$. In this case, $\mathrm{Bin}(n, p)$ approaches the Poisson distribution with intensity parameter μ, leading to nonidentifiability of the parameters (n, p) in the limit. Consequently, more refined estimation techniques become necessary.

Since Fisher (1941), a variety of methods have been proposed to improve upon the sample maximum. A definite answer, however, remains elusive until today. The general lesson from the attempts to obtain better estimators in the small p regime is that further information on n and p is required, which calls for a Bayesian approach. An early Bayesian estimator of the binomial parameters dates back to Draper and Guttman (1971), who suggested to use the posterior mode under a uniform prior for n (upper bounded by some maximal value), and a $\mathrm{Beta}(a, b)$ prior for p with $a, b > 0$. Later, Raftery (1988), Günel and Chilko (1989), Hamedani and Walter (1988), and Berger, Bernardo and Sun (2012), besides others, considered different Bayesian estimators. Raftery (1988), for example, introduced a hierarchical Bayes approach that utilizes a Poisson prior on n with intensity parameter $\lambda > 0$ and a uniform prior distribution on $[0, 1]$ for p. Under the choice $\pi(\lambda) \sim 1/\lambda$ as hyperprior, this hierarchical approach is equivalent to choosing the (improper) scale prior $1/n$ for n. This prior is also recommended as an objective prior for n in Berger, Bernardo and Sun (2012). A broader perspective on objective priors for discrete parameter spaces is offered in Villa and Walker (2014), Villa and Walker (2015), who propose a prior on $n|p$ that depends on the Kullback–Leibler divergence between two successive values of n. The Villa-Walker construction therefore also models a dependency between n and p.

Besides considering the posterior mode and posterior median as estimators, Raftery (1988) suggested to minimize the Bayes risk with respect to the relative quadratic loss. From extensive simulation studies (see the aforementioned references and Section 3 of this article), it is understood that these Bayesian estimators generally deliver good results, especially when compared to frequentist approaches. To the best of our knowledge, however, there is no rigorous theoretical underpinning of these findings. In particular, little 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. First (i), we propose a new class of Bayesian estimators for n, generalizing the approach in Raftery (1988). Second (ii), we analyze the asymptotic behavior of the posterior distribution of n for a large class of priors and asymptotic regimes. This includes statements of posterior consistency as well as a novel Bernstein-von Mises type theorem. Finally (iii), we extend the i.i.d. $\mathrm{Bin}(n, p)$ model to a regression setting and apply the suggested estimators to count the number of fluorophores from super-resolution images. This is a difficult issue of quantitative biology and a target of ongoing research.

Ad (i). We consider product priors of the form $\Pi_n \otimes \Pi_p$ on (n, p) with $\Pi_p \sim \text{Beta}(a, b)$ for some $a, b > 0$ and $\Pi_n(n) \propto n^{-\gamma}$ for all $n \in \mathbb{N}$ and some $\gamma > 1$. Independence of n and p in the prior is a natural assumption and can be justified in our application based on physical considerations (Section 4). The beta prior for p is the standard choice and makes the problem analytically tractable due to its conjugacy property (Draper and Guttman (1971)). The priors $n^{-\gamma}$ for n, which we call scale priors with scaling parameter γ, are widely studied in the literature on the binomial (n, p) problem and its variations, see (Berger, Bernardo and Sun (2012), Link (2013), Raftery (1988), Wang, He and Sun (2007), Tancredi, Steorts and Liseo (2020)).

Based on these prior choices, we focus on two Bayesian scale estimators for n. The first is the posterior mode estimator $\hat{n}_{\text{pm}}$ and the second is the Bayes estimator $\hat{n}_{\text{rql}}$ with respect to the relative quadratic loss, $\ell(x, y) = (x/y - 1)^2$. Following Raftery (1988), the respective estimators are given by

$$\hat{n}_{\text{pm}} = \arg\max_{n \geq M_k} \frac{L_{a,b}(n)}{n^{\gamma}}, \tag{1.2a}$$

$$\hat{n}_{\text{rql}} = \frac{\mathbb{E}[\frac{1}{n}|\mathbf{X}^k]}{\mathbb{E}[\frac{1}{n^2}|\mathbf{X}^k]} = \frac{\sum_{n=M_k}^{\infty} \frac{1}{n^{1+\gamma}} L_{a,b}(n)}{\sum_{n=M_k}^{\infty} \frac{1}{n^{2+\gamma}} L_{a,b}(n)}, \tag{1.2b}$$

where $\mathbf{X}^k = (X_1, \ldots, X_k)$ denotes the data vector and $L_{a,b}(n)$ is the (data dependent) beta-binomial likelihood defined in equation (2.1) below (see also Carroll and Lombard (1985)). In our applications (Section 3 and 4), we assume $\hat{n}_{\text{pm}}$ and $\hat{n}_{\text{rql}}$ to be integer valued by taking the arg max over $\mathbb{N}$ and by rounding $\hat{n}_{\text{rql}}$ to the nearest integer.

Ad (ii). We provide asymptotic conditions under which the marginal posterior for n concentrates all mass around the true population size. As before, we assume product priors on (n, p) with a beta prior on p. For n, we allow general proper priors that decay at most polynomially,

$$\Pi_n(n) \geq \beta n^{-\alpha}, \tag{1.3}$$

for all $n \in \mathbb{N}$ and some $\alpha > 1$ and $\beta > 0$. To investigate the asymptotic behavior of the posterior distribution, we let n and p depend on the sample size k. We formalize this by considering parameter domains of the form

$$\mathcal{M}_k(\lambda) := \left\{(n, p) \colon \frac{1}{\lambda} \leq np \leq \lambda, n \leq \lambda \frac{\sqrt{k}}{\log^6(k)}\right\}, \tag{1.4}$$

where $\lambda > 1$ can be chosen arbitrarily. This class describes binomial variables $\text{Bin}(n, p)$ with expectation values np bounded away from 0 and infinity, such that n grows (at most) slightly slower than $\sqrt{k}$. Under the condition that Π_n satisfies (1.3), posterior contraction around the true population size n_0 is studied in Theorem 1. If n_0 does not grow faster than $k^{1/4}/\log(k)$, we will see that the posterior mass eventually concentrates on the true n_0. In Theorem 2, we then extend our analysis to a different asymptotic domain in which the true population size n_0 stays bounded but p_0 is allowed to decay. Lower bounds that we establish in Theorem 3 guarantee that the rates for consistency in Theorems 1 and 2 are indeed sharp up to logarithmic factors. We also derive a Bernstein-von Mises type result for the posterior on n in Theorem 4, which shows that the limit distribution can be viewed as a discretized version of a normal distribution.

The main building block underlying the recent advances in the frequentist analysis of posterior concentration are the connection to posterior mass conditions and the existence of separating statistical test, see Ghosal, Ghosh and van der Vaart (2000), Ghosal and van der Vaart

(2017), Schwartz (1965). To establish model selection properties of the posterior requires typically different tools (Castillo, Schmidt-Hieber and van der Vaart (2015), Castillo and van der Vaart (2012), Gao, van der Vaart and Zhou (2020)). Since proving that the posterior concentrates on the true population size can be viewed as posterior model selection, it is not surprising that we do not follow the standard posterior contraction proof technique. In fact, a much more refined analysis of the likelihood is necessary and we crucially rely on a decomposition of the log-likelihood via a telescoping sum that is due to Hall (1994). The main challenge in our approach consists of obtaining uniform results over parameter classes where $n \to \infty$ and $p \to 0$ is allowed (in order to capture the small p regime). For fixed n and p, in contrast, posterior consistency as $k \to \infty$ already follows from Doob's consistency theorem (see van der Vaart (1998), Theorem 10.10).

Ad (*iii*). Modern cell microscopy allows researchers to observe the activity and interactions of biomolecules in unprecedented detail. Especially since the development of super-resolution nanoscopy, for which the 2014 Nobel Prize in Chemistry was awarded, it has become an indispensable tool for understanding the biochemical function of proteins (see Hell (2015) for a survey). Super-resolution techniques rely on photon counts obtained from fluorescent markers (or fluorophores), which are tagged to the specific protein of interest and excited by a laser beam. In this article, we are concerned with single marker switching (SMS) microscopy (Betzig et al. (2006), Hess, Girirajan and Mason (2006), Rust, Bates and Zhuang (2006), Fölling et al. (2008)) where the activation of fluorophores and the emission of photons is inherently random: after excitation by a laser, 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, for example, Lee et al. (2012), Rollins et al. (2015), Aspelmeier, Egner and Munk (2015), Staudt et al. (2020). In Section 4 we show how the number of fluorophores can be obtained from a modified binomial (n, p) model. 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 inferred from observations at later time points by linking them through an exponential decay. This leads to a variant of the binomial (n, p) model where the bleaching probability of a fluorophore can be estimated jointly with n_0. We apply this model to experimental data and determine the number of fluorophores on DNA origami test beds.

Outline. This paper is organized as follows. Our results on posterior contraction and the Bernstein-von Mises type theorem can be found in Section 2. For a broader perspective, we also discuss previous results on the asymptotics of several frequentist estimators for n. Section 3 contains an extensive simulation study in which we examine the posterior of n for moderate to large k and compare the finite sample properties of several Bayesian and frequentist estimators. Furthermore, we study the choice of suitable scale priors in different settings and investigate robustness against model deviations from the $\mathrm{Bin}(n, p)$ model. In Section 4, we apply our estimators to data from super-resolution microscopy. The proof of our main posterior contraction result (Theorem 1) is presented in Section 5. Further proofs, auxiliary statements, as well as additional figures are deferred to the Supplementary Material.

2. Asymptotic results. Recall that we observe k independent random variables $X_1, \dots, X_k$ with $\mathrm{Bin}(n, p)$ distribution. We refer to this setting as the binomial (n, p) model. The joint distribution of the data $\mathbf{X}^k = (X_1, \dots, X_k)$ is denoted by $\mathbb{P}_{n,p}$ and the expectation with respect to this distribution is $\mathbb{E}_{n,p}$. We study product priors $\Pi_n \otimes \Pi_p$ on (n, p) and set $\Pi_p = \mathrm{Beta}(a, b)$ with parameters $a, b > 0$. The prior Π_n for n can be chosen as any proper

probability distribution on the positive integers such that condition (1.3) holds for some $\alpha > 1$ and $\beta > 0$. We write $M_k = \max_{i=1,\dots,k} X_i$ for the sample maximum and $S_k = \sum_{i=1}^k X_i$ for the sample sum. The true parameter values are denoted by n_0 and p_0.

For a measurable set $A \subseteq [0,1]$ and $n \in \mathbb{N}$, the joint posterior distribution for (p,n) is given by

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

if $n \geq M_k$ and $\Pi(p \in A, n|\mathbf{X}^k) = 0$ otherwise. The marginal posterior distribution of n is thus

$$\Pi(n|\mathbf{X}^k) \propto \underbrace{\prod_{i=1}^k \binom{n}{X_i} \frac{\Gamma(kn - S_k + b)\Gamma(S_k + a)}{\Gamma(kn + a + b)} \mathbf{1}(n \geq M_k)}_{=:L_{a,b}(n)} \Pi_n(n), \tag{2.1}$$

where Γ is the Gamma function, $\mathbf{1}$ the indicator function, and $L_{a,b}$ the beta-binomial likelihood.

Posterior contraction. Our first result establishes uniform posterior concentration around the true value n_0 over parameters in the set $\mathcal{M}_k(\lambda)$ defined in equation (1.4). Its proof can be found in Section 5.

THEOREM 1. *Consider the binomial (n, p) model under the prior mass condition (1.3). For fixed $\lambda > 1$ and $k \to \infty$,*

$$\sup_{(n_0,p_0)\in\mathcal{M}_k(\lambda)} \mathbb{E}_{n_0,p_0}\left[\Pi\left(n : |n - n_0| \geq \frac{n_0^2 \log^{7/4}(k)}{\sqrt{k}} \middle| \mathbf{X}^k\right)\right] \to 0. \tag{2.2}$$

Equivalently, this result could also be stated in terms of the relative loss $\ell(n, n_0) = |n/n_0 - 1|^2$, which is widely studied in the Bayesian literature for this and related problems, see Smith (1988). A noteworthy consequence of Theorem 1 is that the posterior of n eventually places all mass on the true population size n_0 if the parameters $(n_0, p_0) \in \mathcal{M}_k(\lambda)$ additionally satisfy

$$n_0^2 < \frac{\sqrt{k}}{\log^{7/4}(k)}. \tag{2.3}$$

An inspection of the proof of Theorem 1 reveals that the lower bound on the prior mass condition (1.3) only has to hold for the true value n_0. If we consider sequences of (proper) priors $\Pi_{n,k}$ for n that can change with the sample size k, it can readily be seen from bound (5.14) in the proof that the assertion of the theorem also holds if $\Pi_{n,k}(n) \geq \beta/(nk)^\alpha$ for all positive integers $n \leq \lambda k^{1/2}$ and some $\alpha, \beta > 0$. In particular, it holds for priors with restricted support of the form

$$\Pi_{n,k}(n) \propto f(n)\mathbf{1}\big(n \leq \lambda k^\alpha\big), \tag{2.4}$$

where f satisfies $n^{-\alpha/2} \lesssim f(n) \lesssim n^{\alpha/2}$ for some $\alpha \geq 1/2$.

The techniques used to prove Theorem 1 can also be adapted to asymptotic regimes where n_0 is bounded and p_0 converges to 0 as k tends to infinity. In this case, we depart from the Poisson limit and it should thus become easier to discern the parameters n_0 and p_0. Still, if p_0 approaches zero quickly with increasing k, only a few observations with positive counts will remain, such that the problem becomes difficult again. The next result states that posterior consistency holds in this setting as long as $p_0 \gtrsim \log k/\sqrt{k}$. Its proof is very similar in structure to the one of Theorem 1 and can be found in Section C of the Supplementary Material.

THEOREM 2. *Consider the binomial (n, p) model. For any $B \geq 2$, define the parameter regime*

$$\mathcal{M}_k^{\mathrm{b}}(B) := \left\{(n, p) \colon 2 \leq n \leq B, \frac{\log k}{B\sqrt{k}} \leq p\right\}.$$

If $\Pi_n(n) > 0$ for all $n \in \mathbb{N}$ with $2 \leq n \leq B$, the posterior asymptotically concentrates all mass on the true population size as $k \to \infty$, meaning

$$\sup_{(n_0, p_0) \in \mathcal{M}_k^{\mathrm{b}}(B)} \mathbb{E}_{n_0, p_0}\big[\Pi(n \neq n_0 | \mathbf{X}^k)\big] \to 0.$$

The uniform posterior concentration on the true value n_0 that follows for parameters in the domain $\mathcal{M}_k^{\mathrm{b}}(B)$ (by Theorem 2) and for parameters in $\mathcal{M}_k(\lambda)$ that additionally satisfy (2.3) (by Theorem 1) also implies uniform consistency of the respective posterior mode estimators $\hat{n}_k \in \arg\max_n \Pi(n|\mathbf{X}^k)$. Indeed, for any subset $\mathcal{M}_k$ of the mentioned domains,

$$\sup_{(n_0, p_0) \in \mathcal{M}_k} \mathbb{P}_{n_0, p_0}(\hat{n}_k \neq n_0) \to 0 \tag{2.5}$$

as $k \to \infty$. As a special case, this includes the estimator $\hat{n}_{\mathrm{pm}}$ introduced in equation (1.2a). Furthermore, if $\mathcal{M}_k$ is such that n_0 stays bounded, consistency of the Bayes estimator $\hat{n}_{\mathrm{rql}}$ with respect to the relative quadratic loss given in (1.2b) also follows. The same holds for the Bayesian estimators introduced in Hamedani and Walter (1988) and Günel and Chilko (1989). Since the estimators in Raftery (1988), Berger, Bernardo and Sun (2012), and Link (2013) are based on improper priors for n, our results can be applied to modifications of these estimators where Π_n is restricted to a bounded support.

We next state a lower bound proving that no uniformly consistent estimator for n_0 exists if $n_0/p_0 \gtrsim \sqrt{k}$ (see Section D in the Supplementary Material for a proof). Combined with statement (2.5), this implies that posterior contraction on the true value n_0 is impossible in this regime.

THEOREM 3 (Lower bound). *Let $\eta, \delta > 0$ and fix sequences $(n_k)_k \subset \mathbb{N}$ and $(p_k)_k \subset (0, 1-\delta)$ such that $n_k/p_k \geq \eta\sqrt{k}$ for all k. Define the set $\mathcal{M}_k^* := \{(n_k, p_k), (n_k + 1, p_k')\}$ where $p_k' = \frac{n_k}{n_k+1} p_k$. Then there exists a positive constant $c = c(\eta, \delta)$ such that for any estimator $\hat{n} = \hat{n}(\mathbf{X}^k)$ and all k*

$$\max_{(n_0, p_0) \in \mathcal{M}_k^*} \mathbb{P}_{n_0, p_0}(\hat{n} \neq n_0) \geq c.$$

If the expectation value $n_0 p_0$ is constant or stays bounded away from zero and infinity (and p_0 thus essentially behaves like $1/n_0$), Theorem 3 implies that it is impossible to recover n_0 asymptotically when $n_0 \gtrsim k^{1/4}$. Therefore, the sufficient condition (2.3) for posterior consistency in Theorem 1 is sharp up to logarithmic factors. Similarly, Theorem 3 also implies that the asymptotic recovery of a bounded $n_0 \leq B$ is only possible if $p_0 \gtrsim 1/\sqrt{k}$, which proves that the lower bound on p in Theorem 2 can at most be relaxed by a factor of $\log(k)$. Another consequence of Theorem 3 is that product priors $\Pi = \Pi_n \otimes \Pi_p$ are already asymptotically optimal in the settings of Theorems 1 and 2 (at least up to log-factors). Modeling dependencies between n and p via Π may hence affect the finite sample performance, but it will not improve the asymptotic behavior substantially.

To complete the discussion on posterior concentration, it should be mentioned that another interesting regime occurs if p_0 is bounded away from zero and $n_0 \to \infty$ as $k \to \infty$. Since the sample maximum grows quickly in this case, controlling the posterior requires completely different bounds than before. This regime is of little relevance for our application and we omitted the mathematical analysis in this work. Note that a numerical study in Schneider, Staudt and Munk (2018) indicates that posterior consistency holds in this setting as long as n_0 grows slower than $\sqrt{k}$, which coincides with the lower bound in Theorem 3.

Limiting shape of the posterior. In the regime where the binomial expectation $n_0 p_0$ is bounded away from zero and infinity, we can characterize the limiting distribution of the posterior in the Bernstein-von Mises (BvM) sense. For parametric problems, the standard BvM theorem states, under weak conditions on the prior and the model, that the posterior converges in total variation distance to a normal distribution centered at the MLE (see van der Vaart (1998) for a precise statement). The BvM phenomenon has been studied in a variety of nonstandard settings as well, including estimation of the probability mass function Boucheron and Gassiat (2009), nonregular models Bochkina and Green (2014), and model selection Castillo, Schmidt-Hieber and van der Vaart (2015). To the best of our knowledge, BvM theorems for discrete parameters have not been considered yet. One might wonder in which sense such a limiting shape theorem can hold, since a discrete distribution can not converge to a continuous distribution with respect to the total variation distance.

For the binomial (n, p) problem, we show below that the posterior on n converges in total variation to a discretized version of the normal distribution. The total variation distance between two discrete distributions P and Q defined on the integers is $\mathrm{TV}(P, Q) = \frac{1}{2}\sum_{i\in\mathbb{Z}} |P(i) - Q(i)|$, and we say that an integer-valued random variable X has the discrete normal $\mathcal{N}_{\mathrm{d}}(\mu, \sigma^2)$ distribution if it satisfies $\mathbb{P}(X = j) \propto \exp(-\frac{1}{2\sigma^2}(j-\mu)^2)$ for all $j \in \mathbb{Z}$. This distribution is characterized in Kemp (1997) as the probability distribution on the integers with maximal entropy for given expectation and variance. Its connection to the Jacobi theta functions and other properties are analyzed in Szabłowski (2001).

Asymptotically, the posterior of n will be centered at the estimator

$$\hat{n} := \frac{S_k^2}{S_k^2 - k\sum_{i=1}^k X_i(X_i - 1)} \quad \text{with } S_k = \sum_{i=1}^k X_i. \tag{2.6}$$

In Hoel (1947), this estimator is attributed to Student (1919), who derived it by matching the first two moments of the binomial distribution.

THEOREM 4 (Discrete Bernstein-von Mises). *Suppose that the parameter a in the* Beta(a, b) *prior on p is a nonnegative integer and $\Pi(n) \propto n^{-\alpha}$ for some $\alpha > 1$ and all $n \in \mathbb{N}$. Then, as $k \to \infty$,*

$$\sup_{(n_0,p_0)\in\mathcal{M}_k(\lambda)} \mathbb{E}_{n_0,p_0}\left[\mathrm{TV}\left(\Pi(n = \cdot|\mathbf{X}^k), \mathcal{N}_{\mathrm{d}}\left(\hat{n}, \frac{2n_0^2}{kp_0^2}\right)\right)\right] \to 0.$$

The proof is rather involved and precise bounds for the likelihood ratio in a neighborhood of the true n_0 are required. The main step is to establish that the log-likelihood can locally around n_0 be written as

$$\frac{1}{2}\sum_{i=1}^k (X_i)_2 \log\left(1 - \frac{1}{n}\right) + \frac{S_k^2}{2kn} \tag{2.7}$$

up to terms of negligible order. It can be checked that $n = \hat{n}$ is a maximizer of this expression. A second order Taylor expansion of (2.7) around $\hat{n}$ then shows that the posterior is close to the limit on a localized set. The full proof is deferred to Section E in the Supplementary Material.

Since p_0 is of order $1/n_0$ for parameters (n_0, p_0) in the class $\mathcal{M}_k(\lambda)$, the limit distribution in Theorem 4 converges to the point mass on $\hat{n}$ if $n_0 \ll k^{1/4}$. For $n_0 \gg k^{1/4}$, on the other hand, the limiting variance diverges with k. In this context, we also mention another possibility to define a discretized normal distribution $Z \sim \mathcal{N}_{\mathrm{D}}(\mu, \sigma^2)$ on the integers via $Z := \arg\min_{j\in\mathbb{Z}} |j - X|$ for $X \sim \mathcal{N}(\mu, \sigma^2)$. The distributions $\mathcal{N}_{\mathrm{D}}(\mu, \sigma^2)$ and $\mathcal{N}_{\mathrm{d}}(\mu, \sigma^2)$ are

not the same, but they are close in total variation distance for large σ, see Lemma E.3 in the Supplementary Material. If $n_0 \gg k^{1/4}$, this implies that we can replace the limit distribution $\mathcal{N}_{\mathrm{d}}$ in the BvM type result by $\mathcal{N}_{\mathrm{D}}$.

We conjecture that discretized normal distributions like the ones above will occur as generic posterior limit distributions for a wide range of discrete parameter models, such as the ones considered in Choirat and Seri (2012).

Asymptotic results for frequentist methods. For comparison, we briefly summarize existing asymptotic results for frequentist estimators. Early estimators for n based on the method of moments and the maximum likelihood approach can be found in Haldane (1941) and Blumenthal and Dahiya (1981). In Olkin, Petkau and Zidek (1980), it is shown that these estimators are highly irregular if p is small and methods to stabilize them are proposed. More recently, two further estimators were introduced by DasGupta and Rubin (2005): another modification of the method of moments estimator, and a bias correction of the sample maximum. For the new moments estimator, $\hat{n}_{\mathrm{NME}}$, which depends on the choice of a tuning parameter $\alpha > 0$, it holds that

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

as $k \to \infty$, where n and p are both held fixed. To derive this result, the authors exploit the exponential convergence of the sample maximum to n, which suggests that the limit distribution is only an accurate approximation for very large values of k, especially if p is small. For the bias corrected sample maximum $\hat{n}_{\mathrm{bias}}$, DasGupta and Rubin (2005) derive

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

as $k \to \infty$, where δ_1 denotes the Dirac measure at 1.

The Carroll–Lombard estimator $\hat{n}_{\mathrm{CL}}$ in Carroll and Lombard (1985) is the maximizer of the beta-binomial likelihood in (2.1). It is therefore the posterior mode estimator under a beta prior on p and an improper uniform prior on n. For p constant, $n \to \infty$ and $\sqrt{k}/n \to 0$ as $k \to \infty$, it is known that

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

All of the results above hold for p fixed and hence provide only limited insight into the situation when p is small. A notable extension is discussed in Hall (1994). This article studies a variation $\tilde{n}_{\mathrm{CL}}$ of the Carroll–Lombard estimator by restricting the search for the maximum of the beta-binomial likelihood to a suitable neighborhood around the true n. Since this construction depends on the truth, the maximizer $\tilde{n}_{\mathrm{CL}}$ is in a strict sense not an estimator. It is shown that for $n = n_k \to \infty$ and $p = p_k \to 0$, $np \to \mu \in (0, \infty]$, and $kp^2 \to \infty$,

$$\frac{p\sqrt{k}}{\sqrt{2}}\left(\frac{\tilde{n}_{\mathrm{CL}} - n}{n}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \tag{2.8}$$

as $k \to \infty$. This setup is similar to the one in Theorems 1 and 4, but it does not cover the asymptotic regime considered in Theorem 2. For the asymptotic normality in (2.8), it matters that $\tilde{n}_{\mathrm{CL}}$ is regarded as maximizer over the real numbers and not the integers. To see this, consider a sequence such that $p\sqrt{k}/n \to \infty$. As the rate in (2.8) blows up, we must have that $\tilde{n}_{\mathrm{CL}}$ converges to n in probability, which means that if one replaces $\tilde{n}_{\mathrm{CL}}$ by the closest integer, one recovers the exact value of n with probability increasing to one as $k \to \infty$. Also note that result (2.8) is a specific scenario in a broader context and relies on further technical conditions, like n to be lower bounded by some positive power of k.

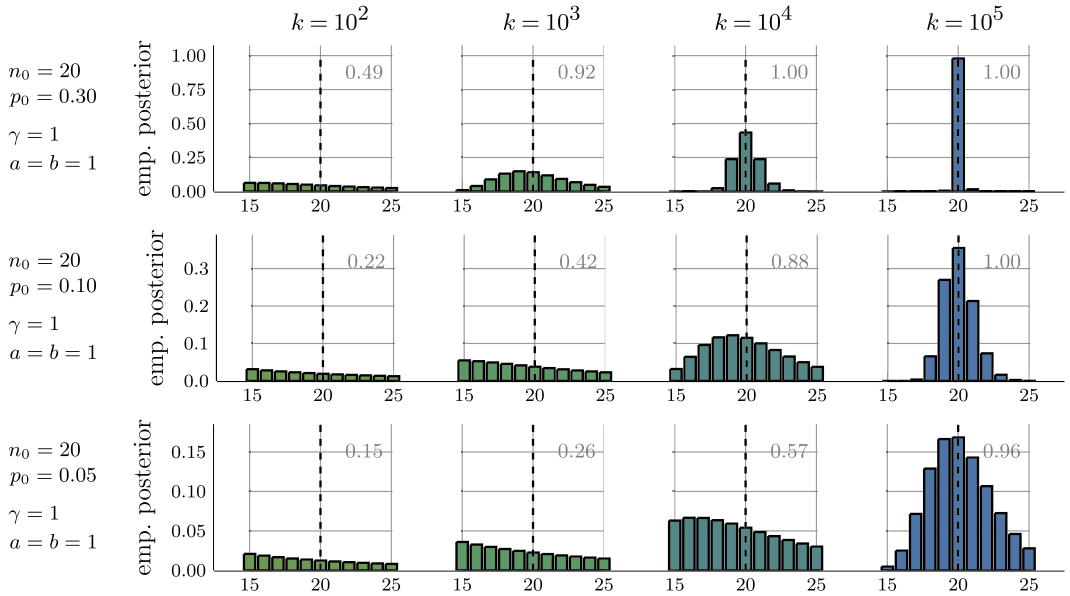

FIG. 1. *Averaged posterior distributions for true parameters* $n_0 = 20$, $p_0 \in \{0.05, 0.1, 0.3\}$, *sample sizes* $100 \le k \le 10^5$, *and* $\gamma = a = b = 1$. *The bar plots display* $\mathbb{E}_{n_0,p_0}[\Pi(n|\mathbf{X}^k)]$ *for different values of* n. *The number in the upper right corner of each graph is the expected posterior mass in the interval* $[15, 25]$.

3. Numerical results. In this section, we numerically investigate the posterior distribution and the finite sample performance of Bayesian estimators for different choices of priors Π_p and Π_n. We consider beta priors with parameters $a, b > 0$ for p, as well as proper and improper scale priors $\Pi_n(n) \sim n^{-\gamma}$ with $\gamma \ge 0$. In situations where we assume a prior guess $\tilde{p}$ for the value of p, the parameters a and b are chosen such that $a \in \{1, 2\}$ and $b = a/\tilde{p} - a$. Then, the Beta(a, b) distribution has expectation $\tilde{p}$ and its probability density function is monotone if $a = 1$, while it is unimodal if $a = 2$. For comparison, we also study the objective prior with $a = b = 1$, corresponding to a uniform distribution on the probability of success p.

Posterior contraction. For $\gamma = a = b = 1$, Figure 1 displays the expected posterior $n \mapsto \mathbb{E}_{n_0,p_0}[\Pi(n|\mathbf{X}^k)]$ for $n_0 = 20$ and $p_0 \in \{0.05, 0.1, 0.3\}$ based on 1000 draws of the data. Figures for different parameters can be found in Section F of the Supplementary Material. For $n_0 = 20$ and $p_0 = 0.3$, Figure 1 demonstrates that the posterior distribution visibly contracts to the true value $n_0 = 20$ for sample sizes $k \ge 10^4$. If $p_0 \le 0.1$, $k = 10^5$ or more observations become necessary for a comparable effect. Figure F.3 in the Supplementary Material shows that increasing n_0 likewise results in broader and less concentrated distributions for given sample sizes k. Changing γ, a, or b has little effect on the shape of the posterior for large values of k, which is in accordance with the Bernstein-von Mises type result in Theorem 4. Still, setting $a = 2$ and $b = 2/p_0 - 2$ notably affects the distributions for $k = 100$ and $k = 1000$ by reducing the bias of the mode, especially when p_0 is small (see Figures F.1 and F.2 in the Supplementary Material).

It is worth pointing out that the posterior of n behaves considerably better than the sample maximum M_k. For example, if $n_0 = 20$ and $p_0 = 0.3$, a sample size of at least $k = 10^{10}$ is needed for $\mathbb{P}_{n_0,p_0}(M_k = 20) \ge 0.35$, while about 10^4 samples are sufficient for $\mathbb{E}_{n_0,p_0}[\Pi(n = 20|\mathbf{X}^k)] \ge 0.35$. If p_0 is set to 0.1 in this comparison, the respective sample sizes are of the dimensions 10^{19} versus 10^5.

Posterior shape. In order to examine the validity of the Bernstein-von Mises type result for finite samples, we compare the posterior of n to the discrete normal $\mathcal{N}_{\mathrm{d}}$ distribution predicted as limit in Theorem 4. Figure 2 depicts several examples of the posterior distribution

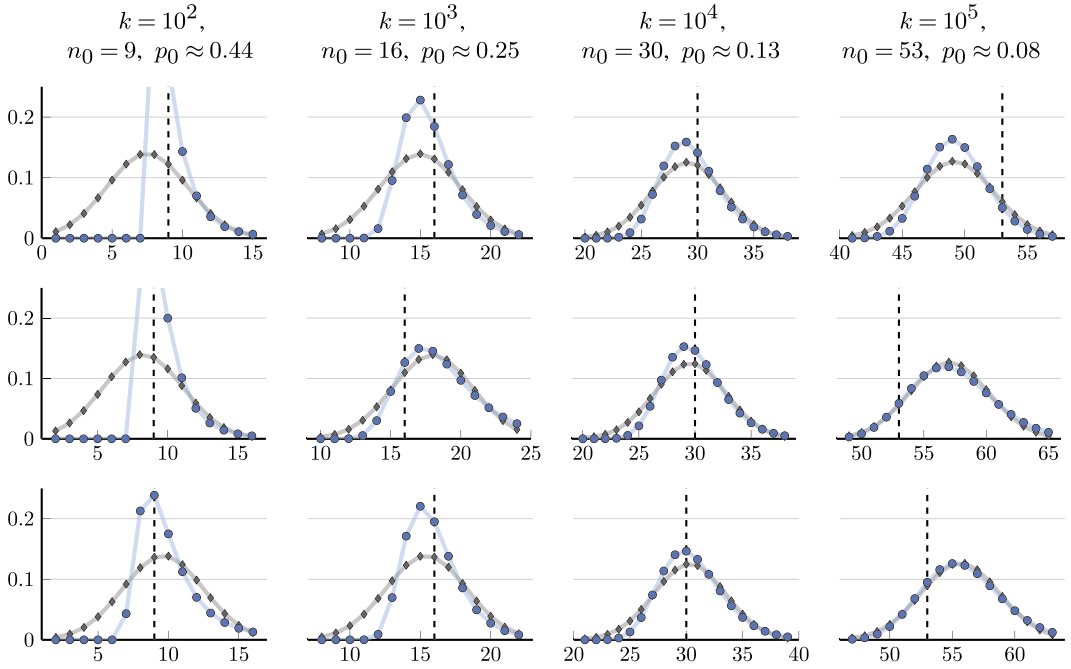

FIG. 2. *Posterior distribution* $\Pi(n = \cdot|\mathbf{X}^k)$ *in blue and discrete normal distribution* $\mathcal{N}_{\mathrm{d}}(\hat{n}, 2n_0^2/kp_0^2)$ *of Theorem 4 in grey. The binomial parameters are chosen according to* $n_0 = \lfloor 3k^{1/4} \rfloor$ *and* $p_0 = 4/n_0$, *where* $\lfloor\cdot\rfloor$ *denotes the floor function. This results in an (asymptotically) constant value* $2n_0^2/kp_0^2 \approx 10$. *The prior parameters are* $\gamma = a = b = 1$. *In each graph, independent realizations of* $\mathbf{X}^k$ *are used. The dashed lines mark the true value* n_0.

in a setting with $n_0 \sim k^{1/4}$ and $p \sim 1/n_0$, such that the variance parameter $\sigma^2 = 2n_0^2/kp_0^2$ of the limiting distribution stays (roughly) constant. While the posterior shape deviates (in part strongly) from the BvM limit for sample sizes $k \leq 10^3$, it clearly approaches the $\mathcal{N}_{\mathrm{d}}$ distribution as k becomes larger. At the same time, the center of the posterior does not seem to concentrate on the true value n_0 as k increases. The posterior often exhibits a less broad distribution than suggested by Theorem 4, especially when the sample maximum M_k reaches into the bulk of the BvM limit for moderate values of k.

Figure 3 shows the total variation distance between the posterior and the BvM limit. This time, we consider settings with $n_0 \sim k^{1/4}$ and $n_0 \sim k^{1/3}$, which are covered by Theorem 4, but

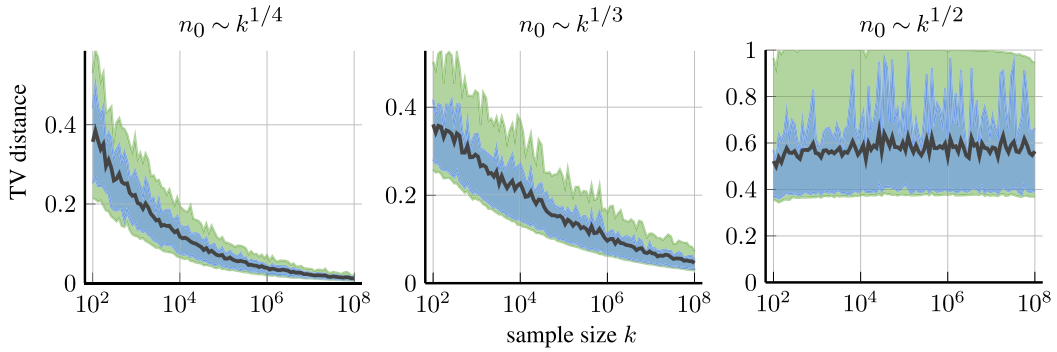

FIG. 3. *Total variation distance between* $\Pi(n = \cdot|\mathbf{X}^k)$ *and* $\mathcal{N}_{\mathrm{d}}(\hat{n}, 2n_0^2/kp_0^2)$ *in dependence of the sample size k for prior parameters* $\gamma = a = b = 1$. *The black line shows the empirical mean over* 100 *independent realizations of* $\mathbf{X}^k$, *while the blue and green areas correspond to the respective* 25–75 *and* 10–90 *percentile ranges. The binomial parameters are chosen as* $n_0 = \lfloor 3k^{1/\delta} \rfloor$ *and* $p_0 = 4/n_0$ *for* $\delta = 4, 3, 2$ *(from left to right).*

also the case $n_0 \sim k^{1/2}$, which falls outside of its scope. One can clearly see the TV distance decreasing in the former two cases, while it does not decay if $n_0 \sim k^{1/2}$. This indicates that the restriction of (n_0, p_0) to $\mathcal{M}_k(\lambda)$ in Theorem 4 cannot be relaxed.

Estimator performance. We next study the finite sample performance of a number of Bayesian and frequentist estimators. In total, the following estimators are considered.

- The scale estimator $\hat{n}_{\mathrm{rql}}$ with respect to the relative quadratic loss defined in (1.2b). It depends on the scale parameter γ and the beta parameters a and b, and we refer to it by SE(γ). Note that the posterior distribution for the scale prior is well defined as long as $a + \gamma > 1$ (see Kahn (1987) for a cautionary note in this context). However, we also report results for SE(0) with $a = 1$, in which case the posterior is no probability distribution, but we still obtain finite estimates when evaluating (1.2b) numerically. The estimator proposed by Raftery (1988) is equivalent to the scale estimator with the choices $\gamma = 1$ and $a = b = 1$, and is denoted by RE in the following.
- The posterior mode estimator $\hat{n}_{\mathrm{pm}}$ defined in (1.2a). We refer to it by PME(γ) and assume the same prior choices as for SE(γ). If $\gamma = 0$, it coincides with the Carroll–Lombard estimator. Furthermore, if $N_0 \in \mathbb{N}$ is chosen sufficiently large, PME(0) in practice also coincides with the estimator proposed by Draper and Guttman (1971), which is the posterior mode estimator under a beta prior on p and $\Pi_n = \mathbf{1}_{\{1,\ldots,N_0\}}$.
- The (frequentist) new moment estimator NME(α) with parameter α, proposed in DasGupta and Rubin (2005). The authors use $\alpha = 1$ in their numerical work.
- The (frequentist) sample maximum MAX.

Note that we do not include the maximum likelihood estimator and the moment estimator (2.6) in our comparison, since their finite sample behavior proved to be very unstable in the range of parameters we consider.

Figure 4 summarizes the performance of the proposed estimators for $n_0 = 20$ and $p_0 = 0.1$ when $k \in \{30, 100, 300\}$. Further simulation results that cover settings with $n_0 \in \{10, 20, 50, 100, 200\}$ and $p_0 \in \{0.05, 0.1, 0.3\}$ can be found in Section F of the Supplementary Material. We observe several salient tendencies among the Bayesian estimators. First, the smaller γ is, the smaller the bias but the larger the variance of the estimates becomes. Estimators with $\gamma = 1$ or 2 typically underestimate n_0, while estimators with $\gamma = 0$ have a larger variability. Second, the bias typically reduces as k is increased from 30 to 300. The variance, on the other hand, only slightly decreases or even increases in some instances. Third, the SE and the PME perform similarly for the same γ, with the former having slightly larger estimates on average. In particular, we can conclude that the posterior mode does not suffer from any peculiar instabilities or other drawbacks. Finally, taking knowledge of p_0 into account (by choosing $a = 2$ and $b = 2/p_0 - 2$) notably reduces the bias of all Bayesian estimators. As expected, this effect is most pronounced for small values of k.

In comparison, the frequentist estimators typically underestimate n_0 more severely than the Bayesian ones. While the NME clearly improves over the sample maximum, it still produces values centered about $n \approx 10$ for both $\alpha = 1$ and $\alpha = 2$ when $n_0 = 20$. We consistently observed that the variance of the NME quickly decreases with increasing k (usually faster than for the Bayesian estimators), but that its bias barely reduces at the same time. Indeed, values estimated by the NME seem to be strongly influenced by the sample maximum, and it seems to inherit the extremely slow convergence toward the real value in the setting of moderate to large k. Similar issues were also observed for the bias reduction estimator proposed by DasGupta and Rubin (2005), which we did not include in our figures. We still stress that the NME(α) with a suitable choice of α is competitive with the Bayesian procedures in some regimes, especially if k is small and p is moderate (see, e.g., Figure F.6 in the Supplementary Material).

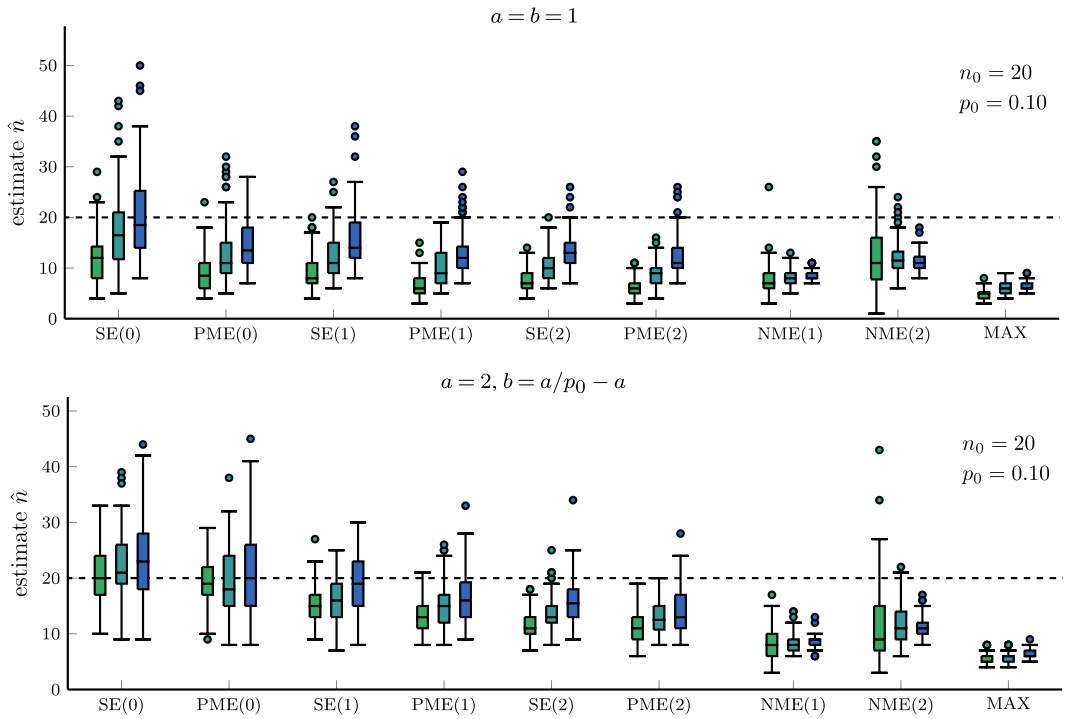

FIG. 4. *Comparison of estimators for n with underlying parameters* $(n_0, p_0) = (20, 0.1)$ *and for three sample sizes* $k = 30$ (*left/green*), $k = 100$ (*middle/turquoise*), $k = 300$ (*right/blue*). *All box plots are based on* 100 *independent repetitions. Outliers are plotted if they deviate from the median by more than* 1.5 *times the interquartile range.*

Prior choices. In the following, we take a systematic look at the influence of the prior choice on the performance of the Bayesian estimators in case of small to moderate sample sizes. Our goal is to establish some practical guidance regarding how to choose γ, a, and b in different scenarios. To this end, we compare the scale estimators SE(γ) with $\gamma \in \{0, 0.5, 1, 2, 3\}$ and the posterior mode estimator PME(0), which corresponds to the Carroll–Lombard estimator, in several simulations, documenting the parameter constellations that perform best.

In a first study, we consider the settings $k \in \{30, 100, 300\}$, $n_0 \in \{20, 50\}$, and $p_0 \in \{0.05, 0.1, 0.3\}$, while assuming a good guess $\tilde{p} = p_0$ that correctly informs the Beta(a, b) prior on p via $a = 2$ and $b = a/\tilde{p} - a$, such that its expectation is $\tilde{p}$. For all pairs (n_0, p_0) and each estimator $\hat{n}$, we empirically approximate:

- the relative mean squared error (RMSE) given by $\mathbb{E}_{n_0, p_0}[(\hat{n}/n_0 - 1)^2]$,
- the bias $\mathbb{E}_{n_0, p_0}[\hat{n}] - n_0$ of the estimator,

by averaging over 1000 realizations of $\mathbf{X}^k$. In Table 1, we present the estimators that have the lowest RMSE and the lowest bias for the different choices of k. The outcome generally advises to select smaller values of γ the smaller p_0 is expected to be. We only found minor differences between the PME(0) and the SE(0). Both of them outperform the other estimators in the regime of very small p_0. The drawback of these estimators is their high variance, which is why larger choices of γ become preferable for low RMSEs as k increases. The similarity of Table 1(a) and 1(b) for $n_0 = 20$ and $n_0 = 50$ suggests that the influence of n_0 is weaker than the one of p_0 for the optimal estimator choice.

Our next study covers a setting that is motivated by the data example in Section 4, and we select $n_0 = 15$, $p_0 = 0.0339$, and $k = 94$. This time, our focus lies on the influence of

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

(a) $n_0 = 20$				(b) $n_0 = 50$			
p_0	k	RMSE	bias	p_0	k	RMSE	bias
0.05	30	PME(0)	SE(0)	0.05	30	PME(0)	SE(0)
0.05	100	PME(0)	SE(0)	0.05	100	PME(0)	SE(0)
0.05	300	SE(0.5)	PME(0)	0.05	300	SE(0.5)	PME(0)
0.1	30	PME(0)	SE(0)	0.1	30	PME(0)	SE(0)
0.1	100	SE(0.5)	PME(0)	0.1	100	SE(0.5)	PME(0)
0.1	300	SE(1)	PME(0)	0.1	300	SE(1)	SE(0.5)
0.3	30	SE(2)	SE(1)	0.3	30	SE(1)	SE(0.5)
0.3	100	SE(3)	PME(0)	0.3	100	SE(3)	PME(0)
0.3	300	SE(3)	SE(2)	0.3	300	SE(3)	PME(0)

the beta prior parameters $a \in \{1, 2\}$ and $b = a/\tilde{p} - a$. We consider four different scenarios: no information about p_0 (setting $\tilde{p} = 0.5$), accurate information ($\tilde{p} = p_0$), underestimation ($\tilde{p} = 0.5p_0$), and overestimation ($\tilde{p} = 1.5p_0$).

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 a good guess for p_0 is available. 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 instabilities and substantial overestimation of n_0 if γ is small. Here, estimators with (proper) prior choices $\gamma = 1$ and 2 perform very well: the tendency of overestimation caused by the choice $\tilde{p} = 0.5p_0$ is in part compensated by the tendency of underestimation due to the higher value of γ.

Overall, our findings confirm that the smaller p_0, the more difficult it becomes to estimate n_0 and the smaller γ should be picked. A smaller γ, however, increases the variance of the posterior distribution and leads to estimators that are potentially more sensitive against miss-specification in the beta prior. This is further investigated in Table 3, where we compare the sensitivity of estimators corresponding to $\gamma = 0$ and $\gamma = 1$. Miss-specifying $\tilde{p} = 0.5p_0$ leads to severe overestimates $\mathbb{E}_{n_0,p_0}[\hat{n}] \approx 2n_0$ for PME(0), while SE(1) is less sensitive in this regard. Selecting $\gamma = 0$ can therefore help to estimate n_0 in very difficult scenarios, but it can also lead to heavily biased results if $\tilde{p}$ is chosen too small.

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

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

$\tilde{p}$	a	est.	RMSE	bias	$\tilde{p}$	a	est.	RMSE	bias
0.5	1	SE(0.5)	0.478	−10.17	$1.5p_0$	1	SE(0)	0.12	−3.73
	1	SE(0)	0.395	−9		2	SE(0)	0.121	−4.69
p_0	2	PME(0)	0.034	−0.266	$0.5p_0$	1	SE(1)	0.036	−0.032
	2	SE(0)	0.036	−0.043		2	SE(2)	0.025	−0.55

TABLE 3
Sensitivity of SE(1) *and* PME(0) *against miss-specification of* $\tilde{p}$. *The value* a *is set to* 2, *all other parameters are selected as in Table* 2. *The behavior of* PME(0) *and* SE(0) *is comparable in this setting*

Estimator	$\tilde{p}$	RMSE	bias
SE(1)	p_0	0.122	−4.85
	$0.5p_0$	0.129	4.43
	$1.5p_0$	0.279	−7.73
PME(0)	p_0	0.034	−0.27
	$0.5p_0$	1.002	14.32
	$1.5p_0$	0.139	−5.09

To study this question numerically, we generated 1000 data sets X_i, $i = 1, \dots, k$, with sample size $k = 100$, where each observation X_i is drawn independently from a $\mathrm{Bin}(n_i, p_0)$ distribution and each n_i is a realization of a binomial random variable $N \sim \mathrm{Bin}(\tilde{n}, \tilde{p})$. For each sample, p_0 is drawn from a 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 constant n_0 (chosen as the integer nearest to $\mathbb{E}[N] = \tilde{n}\tilde{p}$) and the same realization of p_0. For both scenarios, we simulated the RMSE with respect to n_0 and record their ratios in Table 4 for parameters $\tilde{n}$ and $\tilde{p}$ resembling the data example in Section 4. The resulting ratios are all close to one, which suggests a stable performance of the estimators: estimating n_0 from a sample with heterogeneous n_i (randomly drawn from N) instead of constant n_0 (close to $\mathbb{E}[N]$) does not affect the RMSE much (on average).

4. Data example. We now apply the previously described Bayesian estimators to quantify the number of fluorescent molecules in super-resolution microscopy. Reliable methods for this task are highly relevant in quantitative cell biology, which aims to determine the concentration of specific biomolecules, like proteins, in the cell. For general information, see Lee et al. (2012), Rollins et al. (2015), Ta et al. (2015), Aspelmeier, Egner and Munk (2015), Karathanasis et al. (2017), Staudt et al. (2020), and references therein.

Super-resolution microscopy. The term super-resolution microscopy denotes a family of recently developed techniques of fluorescence microscopy. It describes the ability to achieve resolutions below the diffraction limit of visible light (about 250–500nm), which limits classical modes of optical microscopy (Hell (2009)). The central idea is to separate photon emissions of spatially close fluorescent markers (fluorophores) in time, for example, by making

TABLE 4
Ratios of the RMSE for i.i.d. and non-i.i.d. samples (*RMSE-R*) *for the estimators* SE(γ), PME(0), *and the Raftery estimator* RE. *The beta prior for* SE *and* DGE *uses* $a = 2$ *and* $b = 38$

	$\tilde{n} = 8$	$\tilde{n} = 22$
Estimator	RMSE-R	RMSE-R
SE(0.5)	1.022	1.130
SE(1)	1.011	1.067
SE(2)	1.020	1.010
PME(0)	1.032	1.073
RE	0.988	0.981

them switch between active and inactive states (until they bleach and become permanently inactive). In practice, the separation in time is realized by applying an excitation laser with low intensity, such that only a small fraction of fluorophores in the sample are in the active state during a given frame of observation. By combining the resulting "sparse" information recorded over a series of frames, an increased resolution of up to 20–30nm can be achieved. See Betzig et al. (2006), Rust, Bates and Zhuang (2006), Hess, Girirajan and Mason (2006), or Fölling et al. (2008) for different variants of this principle.

Experimental setup. Our data has been recorded at the Institute for Nanophotonics Göttingen e.V. In a preparational step, DNA origami molecules (Schmied et al. (2014)) were dispersed on a microscopic cover slip. DNA origami are nucleotide sequences engineered in such a way that they fold into a desired shape and that fluorophores can attach to them (see Figure 5(a)). In the experiment, Alexa647 fluorophores with 22 different types of anchors were used, each matching a different anchor spot on the origami. The attachment process itself is random and is expected to occur with a probability between 0.6 and 0.75 according to the manufacturer. Hence, about 13 to 17 fluorophores should on average be attached to a single DNA origami.

The experiment was initialized in such a way that most fluorophores occupy their active state in the first frame. All origami are therefore visible as bright spots in Figure 5(b). Note that individual fluorophores occupying the same origami can not be discerned in this image; this becomes possible only by analyzing later frames where most fluorophores are inactive and markers show up individually (see the supplementary video). Each frame had an exposure time of 15ms, and 14,060 consecutive frames were recorded in total over a time span of about 3.5 minutes.

Counting fluorophores. 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 active state within the respective origami. Thus, an origami is invisible if all of its fluorophores are inactive, but its location on the image is still known from the first frame. This allows us to register 94 regions of

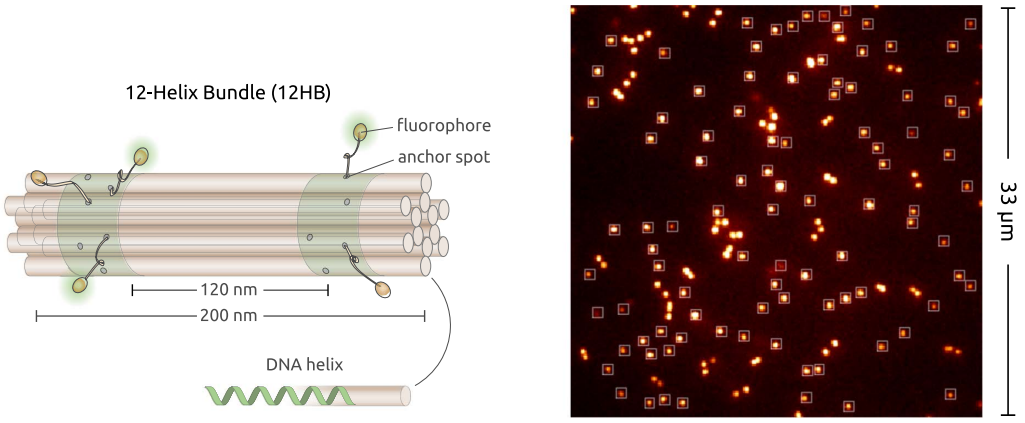

FIG. 5. (*a*) *Schematic drawing of the DNA origami used in the experiment. The origami is a tube-like structure that consists of* 12 *suitably folded DNA helices. In each of the two highlighted green regions up to* 11 *fluorescence markers can anchor.* (*b*) *First frame of the sequence of microscopic images. The* 94 *regions of interest* (*ROIs*) *that were chosen for analysis are identified by white boxes. No overlap 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.*

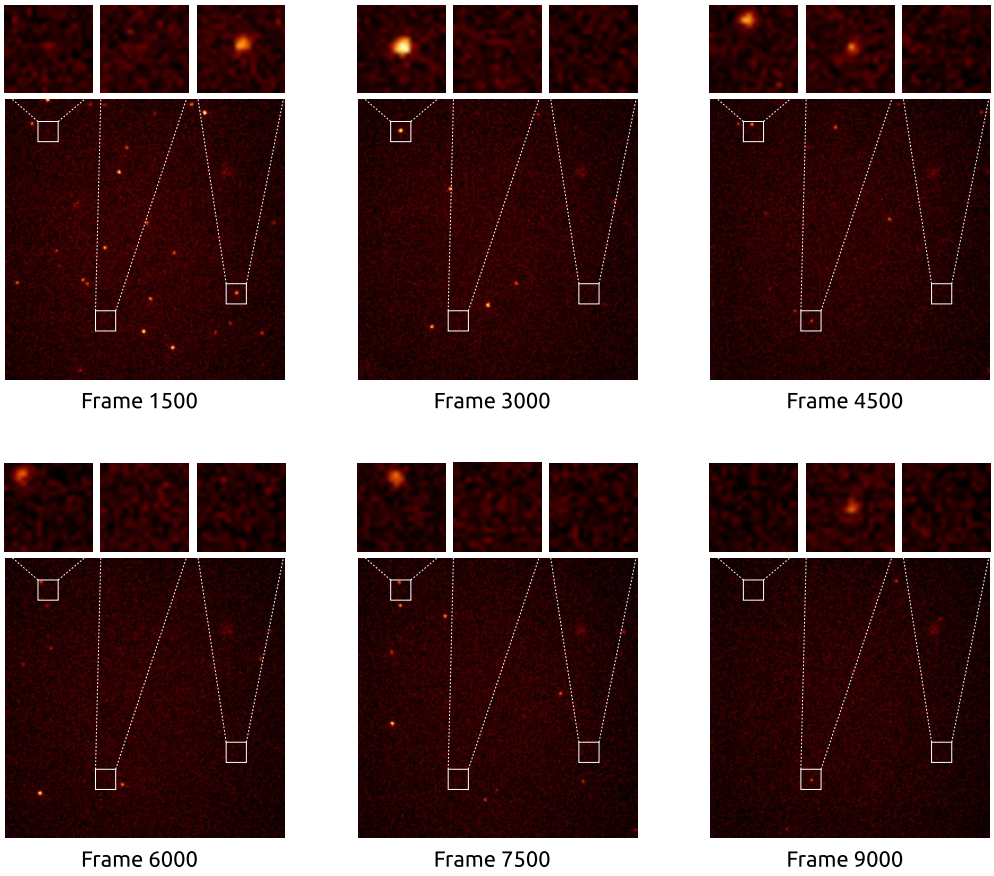

FIG. 6. *Six selected frames from the data set of recorded origami. The (physical) time difference between two consecutive images in this figure is roughly* 23 *seconds. Bleaching causes the number of visible origami to decrease with increasing frame number, while switching causes that unbleached origami are visible only in some frames.*

interest (ROIs) marked in Figure 5(b). For illustration, six microscopic frames recorded at the times $t \in \{1500, 3000, 4500, 6000, 7500, 9000\}$ are visualized in Figure 6. The influence of switching and bleaching on the observations is clearly visible.

We aim to estimate the number of fluorophores attached to each origami, which is expected to be between 13 and 17. For simplicity, we assume 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

$$n_t = n_0(1 - p_b)^t, \tag{4.1}$$

where p_b denotes the bleaching probability. The brightness observed for a spot in frame t is proportional to the (random) number X_t of active fluorophores during the frame's exposure. This number is binomially distributed, $X_t \sim \mathrm{Bin}(n_t, p)$, where p denotes the (time-independent) probability that an unbleached fluorophore is in its active state. We will estimate n_0 and p_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 prior information from a similar experiment where each origami has been designed to carry exactly one fluorophore. We calculate the average ratio between the number of frames where the fluorophore is active (a bright spot is seen) and the total number of frames before bleaching, which yields $\tilde{p} \approx 0.0339$ as a prior guess for p. Therefore, we are indeed in the difficult small-p regime of the binomial

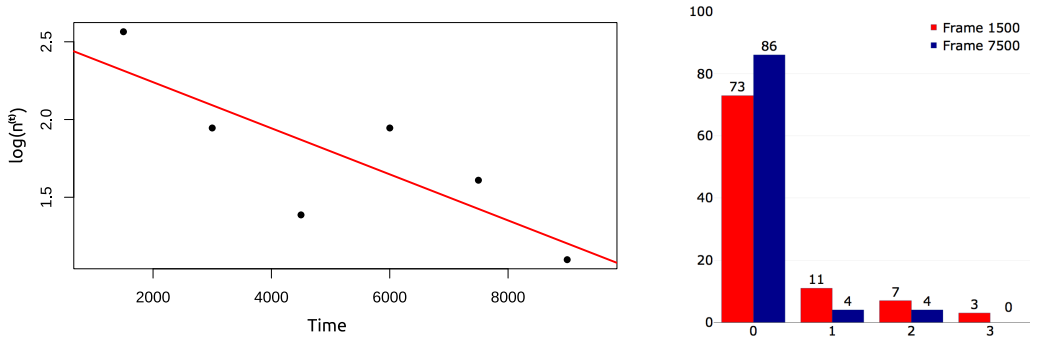

FIG. 7. *(a) Log-linear fit described by $n_t = n_0(1-p_b)^t$ for the SE with $\gamma = 0.5$. (b) Bar charts of the observed numbers of fluorophore molecules for time frames* 1500 *and* 7500.

(n, p) problem and will estimate n_t via the Bayesian scale estimators (1.2), using the notation (SE, PME) of Section 3. The beta prior for SE and PME uses the parameters $a = 2$ and $b = 2/\tilde{p} - 2 \approx 56.99$. We choose the unimodal prior with $a = 2$, as suggested by Table 2, since we assume that our guess $\tilde{p}$ is reasonably accurate. Note that a finer degree of modeling would require to view n_0, n_t and p as random variables instead of constants. However, as shown at the end of 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 active 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, about 23 seconds into the experiment. To mitigate the influence of correlations between observations (since X_t and X_{t+1} for a spot can not 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 $t \in \{1500, 3000, 4500, 6000, 7500, 9000\}$ depicted in Figure 6. 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 typically at most one fluorophore of each origami is active.

The results for the scale estimator SE(0.5) are depicted in Figure 7(a), which shows the log-linear fit for model (4.1). The point estimates of n_0 and p_b for different estimators are summarized in Table 5. Given that the true n_0 in this experiment is expected to be between 13 and 17, we can see that the scale estimators with an improper prior ($\gamma \leq 1$) produce the

TABLE 5
Estimates of the bleaching probability p_b and the number n_0 of fluorophore molecules on single DNA origami

Estimator	n_0	$p_b \cdot 10^3$
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
PME(0)	**16**	**0.167**

most reasonable results. This is in agreement with our observations in Section 3, where we noted that priors putting a lot of weight on large values of n perform better for small p by correcting for the inherent tendency to underestimate (see Table 2). To illustrate the difficulty of this problem, Figure 7(b) shows exemplary counting results for $t \in \{1500, 7500\}$. Note that estimates for each n_t are exclusively based on observations $X_t \leq 3$, where a great majority is even zero.

5. Proof of Theorem 1. In the following, we prove the posterior contraction result of Theorem 1. Since this involves numerous steps, we begin by an outline of the main ideas. Auxiliary results needed by the proof are collected in Section B of the Supplementary Material.

Outline. Throughout the proof, we fix some $\lambda > 1$ and consider a generic sequence $(n_k, p_k)_k$ of parameters that satisfies $(n_k, p_k) \in \mathcal{M}_k(\lambda)$ for all $k \in \mathbb{N}$ with $\mathcal{M}_k(\lambda)$ as defined in (1.4). Since the convergence in Theorem 1 is uniform over $\mathcal{M}_k(\lambda)$, we emphasize that our arguments are indeed independent of the specific choice of $(n_k, p_k)_k$ and all bounds are controlled by the parameter λ alone. For brevity, we usually write $\mathbb{P}_k$ and $\mathbb{E}_k$ instead of $\mathbb{P}_{n_k,p_k}$ and $\mathbb{E}_{n_k,p_k}$ from now on.

Let $A_k \subset \mathbb{N}$ be a series of sets that do not contain the true parameter value n_k. The first step of the proof consists of bounding the (marginal) posterior probability $\Pi(n \in A_k|\mathbf{X}^k)$ in terms of fractions $L_{a,b}(n)/L_{a,b}(n_k)$ of beta-binomial likelihoods (defined in (2.1)) for integers $n \in A_k$. Recall that M_k denotes the sample maximum and $S_k = \sum_{i=1}^k X_i$ the sample sum. Consider the function $R : [0, \infty) \times (0, \infty) \times [M_k, \infty) \to (0, \infty)$,

$$R(a, b, m) = \prod_{i=1}^{k} \frac{\Gamma(m+1)}{\Gamma(X_i+1)\Gamma(m-X_i+1)} \frac{\Gamma(km - S_k + b)\Gamma(S_k + a)}{\Gamma(km + a + b)}, \tag{5.1}$$

which is well defined (even for $a = 0$) if $S_k > 0$. In particular, note that $R(a, b, n) = L_{a,b}(n)$ for $a, b \in (0, \infty)$ and $n \geq M_k$, so that one can write

$$\frac{R(a, b, n)}{R(a, b, n_k)} = \exp\Big(k \int_{n_k}^{n} f'(m)\, \mathrm{d}m\Big), \tag{5.2}$$

where $f(m) := \frac{1}{k} \log R(a, b, m)$ is differentiable. The derivative $f'(m)$ is studied in Hall (1994).

The remainder of the proof focuses on bounding $f'(m)$. This includes the definition of an event $\mathcal{X}_k$ that satisfies $\mathbb{P}_k(\mathcal{X}_k) \to 1$ for $k \to \infty$. We construct this event in such a way that M_k, S_k, and the factorial moments $(X_i)_j$, where $(c)_j := c \cdot (c-1) \cdots (c-j+1)$ for $c \in \mathbb{R}$ and $j \in \mathbb{N}$, exhibit benign properties if $\mathbf{X}^k \in \mathcal{X}_k$. We also need to distinguish between the cases $m \leq n_k$ and $m > n_k$, for which we have to lower-, respectively upper-bound $f'(m)$ on $\mathcal{X}_k$. This requires several technical interim steps, which are largely outsourced to Section B in the Supplementary Material. Combining the resulting bounds yields an upper bound for $\Pi(n \in A_k|\mathbf{X}^k)$ that can be used to show consistency in the asymptotic setting explored in Theorem 1 if the sets A_k are chosen suitably.

PROOF OF THEOREM 1. Let $A_k \subset \mathbb{N}$ be sets such that $n_k \notin A_k$ for all k. It will later become evident how these sets are best be chosen. First, observe that

$$\Pi\big(n \in A_k|\mathbf{X}^k\big) = \frac{\sum_{n \in A_k, n \geq M_k} L_{a,b}(n)\Pi_n(n)}{\sum_{n=M_k}^{\infty} L_{a,b}(n)\Pi_n(n)} \leq \sum_{n \in A_k, n \geq M_k} \frac{L_{a,b}(n)\Pi_n(n)}{L_{a,b}(n_k)\Pi_n(n_k)}.$$

Under the assumption that $S_k \geq 2$, which we justify below, we can apply Lemma B.3 and find

$$\frac{L_{a,b}(n)}{L_{a,b}(n_k)} \leq \frac{R(\lfloor a \rfloor, b, n)}{R(\lceil a \rceil, b, n_k)} \leq c_1 \frac{kn_k}{S_k} \frac{R(\lfloor a \rfloor, b, n)}{R(\lfloor a \rfloor, b, n_k)}$$

for $c_1 = 2(1 + \lceil a \rceil + b)$, where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ denote the ceiling and floor functions, and where R was defined in (5.1). It follows that

$$\Pi\big(n \in A_k | \mathbf{X}^k\big) \leq c_1 \frac{kn_k}{S_k} \sum_{n \in A_k, n \geq M_k} \exp\Big(k \int_{n_k}^{n} f'(m)\,\mathrm{d}m\Big) \frac{\Pi_n(n)}{\Pi_n(n_k)}, \tag{5.3}$$

where $f(m) = \frac{1}{k} \log R(\lfloor a \rfloor, b, m)$. In case that $n < n_k$, we find $\int_{n_k}^{n} f'(m)\,\mathrm{d}m = -\int_{n}^{n_k} f'(m)\,\mathrm{d}m$. For an upper bound on the posterior we thus need a lower bound of $f'(m)$ if $M_k \leq m \leq n_k$ and an upper bound if $m \geq n_k$. Since f only depends on a via $\lfloor a \rfloor$, we for brevity write $a \in \mathbb{N}_0$ to denote $\lfloor a \rfloor$ from now on. Lemma 4.1 in Hall (1994) states that

$$\sum_{j=1}^{r} \frac{1}{c - j + 1} = \sum_{j=1}^{r} \frac{(r)_j}{(c_j)_j\, j}$$

for integers $r \in \mathbb{N}$ and positive numbers $c > k - 1$. We therefore find

$$\begin{aligned} f'(m) &= \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{X_i} \frac{1}{m - j + 1} - \sum_{j=1}^{S_k + a} \frac{1}{km + a + b - j} \\ &= \sum_{j=1}^{M_k} \frac{T_j - U_j}{j} - \sum_{j=M_k+1}^{S_k+a} \frac{U_j}{j} \end{aligned} \tag{5.4}$$

with

$$T_j := \frac{1}{k} \sum_{i=1}^{k} \frac{(X_i)_j}{(m)_j} \quad \text{and} \quad U_j := \frac{(S_k + a)_j}{(km + a + b - 1)_j} \tag{5.5}$$

for $j \leq M_k$ and $j \leq S_k + a$, respectively. If $j > M_k$, we define $T_j := 0$ for all $j > M_k$. The expectation of T_j is given by $t_j := \mathbb{E}_k[T_j] = (n_k)_j (p_k)^j / (m)_j$, which follows from

$$\mathbb{E}_k\big[(X_i)_j\big] = \sum_{x=j}^{n_k} \binom{n_k}{x} (x)_j\, p_k^x q_k^{n_k - x} = (n_k)_j (p_k)^j \underbrace{\sum_{y=0}^{n_k - j} \binom{n_k - j}{y} p_k^y q_k^{n_k - j - y}}_{=1}$$

for all $i = 1, \ldots, k$, where we set $q_k := 1 - p_k$ and substituted $y = x - j$.

Next, recall that $\lambda > 1$ is the constant in the definition of the spaces $\mathcal{M}_k(\lambda)$. For a fixed positive and diverging sequence $l_k = o(\sqrt{\log(k)})$ and $c_2 = 2\lambda(\lambda + 1)$, we introduce the events

$$\begin{aligned} \mathcal{R}_k &:= \big\{\min(n_k, l_k) \leq M_k \leq 2\log(k)\big\}, \\ \mathcal{T}_k &:= \bigcap_{j=1}^{M_k} \big\{(m)_j | T_j - t_j | \leq \sqrt{(c_2 j)^j l_k \log(k)/k}\big\}, \\ \mathcal{S}_k &:= \big\{|S_k - kn_k p_k| \leq \sqrt{\lambda k \log(k)}\big\}, \end{aligned} \tag{5.6}$$

and denote their intersection $\mathcal{R}_k \cap \mathcal{T}_k \cap \mathcal{S}_k$ by $\mathcal{X}_k$. The probability of the event $\mathcal{T}_k$ is independent of m due to the definition of T_j. On the event $\mathcal{S}_k$, Lemma B.5 grants us the additional property

$$|U_j - u_j| \leq j \sqrt{\frac{\lambda \log(k)}{k}} \Big(\frac{c_3}{m}\Big)^j \quad \text{with } u_j := \frac{(kn_k p_k + a)_j}{(km + a + b - 1)_j}$$

for $j \le S_k + a$ and $c_3 = 2e^2(3\lambda + a + 1)$. If $k/\log(k) \ge 4\lambda^3$, then $k/2\lambda \le S_k \le 2\lambda k$ and $S_k \ge 2$ on $\mathcal{S}_k$. Hence, equations (5.3) and (5.4) apply on $\mathcal{X}_k$ if k is sufficiently large. Also, we can use

$$c_1 k n_k / S_k \le 2\lambda c_1 n_k \tag{5.7}$$

to bound the factor preceding the sum in (5.3).

For the remainder of the proof, we can restrict $\mathbf{X}^k$ to $\mathcal{X}_k$ since

$$\mathbb{E}_k\big[\Pi\big(n \ne n_k|\mathbf{X}^k\big)\big] - \mathbb{E}_k\big[\mathbf{1}_{\mathcal{X}_k}\Pi\big(n \ne n_k|\mathbf{X}^k\big)\big] \le \mathbb{P}_k\big(\mathcal{X}_k^{\mathrm{c}}\big) \longrightarrow 0 \tag{5.8}$$

uniformly over $\mathcal{M}_k(\lambda)$ for $k \to \infty$. To show this, we bound

$$\mathbb{P}_k\big(\mathcal{X}_k^{\mathrm{c}}\big) \le \mathbb{P}_k\big(\mathcal{S}_k^{\mathrm{c}}\big) + 2\mathbb{P}_k\big(\mathcal{R}_k^{\mathrm{c}}\big) + \mathbb{P}_k\big(\mathcal{T}_k^{\mathrm{c}} \cap \mathcal{R}_k\big).$$

The first contribution vanishes by the application of Chebyshev's inequality (see, e.g., DeGroot and Schervish (2012)), because $\mathbb{E}_k[S_k] = kn_k p_k$ and $\operatorname{Var}_k[S_k] = kn_k p_k(1-p_k) \le k\lambda$. The second term is controlled by Lemma B.2. For the last term, observe that

$$\operatorname{Var}\big[(X_i)_j\big] \le \big(2jnp(np+1)\big)^j \le \big(2j\lambda(\lambda+1)\big)^j = (c_2 j)^j$$

by Lemma B.1. For any $r > 0$, Chebyshev's inequality yields

$$\mathbb{P}_k\Bigg(\underbrace{\bigg|\frac{1}{k}\sum_{i=1}^{k}(X_i)_j - \mathbb{E}\big[(X_1)_j\big]\bigg| > \sqrt{\frac{r(c_2 j)^j}{k}}}_{=:\mathcal{T}_{jk}^{\mathrm{c}}(r)}\Bigg) \le \frac{\operatorname{Var}[(X_1)_j]/k}{r(c_2 j)^j/k} \le \frac{1}{r}.$$

With $r = l_k \log(k)$ and $M_k \le 2\log(k)$ on $\mathcal{R}_k$,

$$\mathbb{P}_k\big(\mathcal{T}_k^{\mathrm{c}} \cap \mathcal{R}_k\big) = \mathbb{P}_k\Bigg(\bigcup_{1 \le j \le 2\log(k)} \mathcal{T}_{jk}^{\mathrm{c}}\big(l_k \log(k)\big)\Bigg) \le \frac{2\log(k)}{l_k \log(k)} \longrightarrow 0$$

follows. It is important to note that the upper bounds in these inequalities are all controlled by λ, which implies that the convergence in (5.8) is indeed uniform over $\mathcal{M}_k(\lambda)$.

Auxiliary lower bound. For $M_k \le m < n_k$, we prove a lower bound for $f'(m)$. We may assume that $M_k \ge l_k \to \infty$ for $k \to \infty$ in this case, since $\mathbf{X}^k \in \mathcal{R}_k$. For k such that $l_k \ge 4$, equation (5.4) yields

$$f'(m) \ge \sum_{j=1}^{4} \frac{T_j - U_j}{j} - \sum_{5}^{S_k+a} \frac{U_j}{j}, \tag{5.9}$$

as $T_j \ge 0$ for all j. Due to the definition of $\mathcal{M}_k(\lambda)$, we can (generously) bound $m < n_k \le \lambda\sqrt{k\log(k)}$ and

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

To handle the terms in (5.9) with $j \ge 2$, we exploit that $\mathbf{X}^k \in \mathcal{T}_k$ and apply $(m)_j \ge (m/e^2)^j$ (see Lemma B.4) in order to derive

$$\sum_{j=2}^{4} \frac{|T_j - t_j|}{j} \le \sqrt{\frac{l_k \log(k)}{k}} \sum_{j=2}^{4} \bigg(\frac{\sqrt{c_2 j}}{m/e^2}\bigg)^j \le 2\frac{4c_2 e^4}{m^2}\sqrt{\frac{l_k \log(k)}{k}}$$

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

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

for $m \ge l_k \ge 2c_3$. By applying Lemma B.6 with $c_4 = 6e^2(\lambda + a)$ and using $S_k \le 2k\lambda$ on $\mathcal{S}_k$, we furthermore observe

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

for all k (and thus m) that are sufficiently large. The first result of Lemma B.7 combined with $m < n_k \le \lambda\sqrt{k\log(k)}$ reveals

$$\sum_{j=2}^{4} \frac{t_j - u_j}{j} \ge \frac{1}{2\lambda^2}\frac{n_k - m}{n_k m^3} - \frac{3c_5^4}{mk} \ge \frac{1}{2\lambda^2}\frac{n_k - m}{n_k m^3} - 2\frac{2\lambda c_5^4}{m^2}\sqrt{\frac{\log(k)}{k}},$$

where $c_5 = 3\lambda(1 + a + b) + 2a + 4$. All bounds calculated above can be inserted into inequality (5.9), yielding

$$\begin{aligned} f'(m) &\ge (T_1 - U_1) + \sum_{j=2}^{4} \frac{T_j - t_j}{j} + \sum_{j=2}^{S_k+a} \frac{u_j - U_j}{j} + \sum_{j=2}^{4} \frac{t_j - u_j}{j} - \sum_{j=5}^{S_k+a} \frac{u_j}{j} \\ &\ge C_1 \frac{n_k - m}{n_k m^3} - \frac{C_2}{m^2}\sqrt{\frac{l_k \log(k)}{k}} + \underbrace{\left[\frac{1}{4\lambda^2}\frac{n_k - m}{n_k m^3} - 2\left(\frac{c_4}{m}\right)^5\right]}_{=:h(m)} \end{aligned} \tag{5.10}$$

with $C_1 = 1/(4\lambda^2)$ and $C_2 = 2(\lambda(a+1) + 4c_2e^4 + \sqrt{\lambda}c_3^2 + 2\lambda c_5^4)$.

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

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

from equation (5.4). For $j = 1$,

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

where we used that $S_k \le 2\lambda k$ on the event $\mathcal{S}_k$. Next, we set $m_0 := 4c_2e^4$ and derive

$$\begin{aligned} \sum_{j=2}^{M_k} \frac{|T_j - t_j|}{j} &\le \sqrt{\frac{l_k \log(k)}{k}} \sum_{j=2}^{M_k} \left(\frac{\sqrt{c_2 j}}{m/e^2}\right)^j \\ &\le \frac{c_2 M_k e^4}{m^2}\sqrt{\frac{l_k \log(k)}{k}} \sum_{j=0}^{\lfloor m \rfloor} \left(\frac{e^2\sqrt{c_2}}{\sqrt{m}}\right)^j \\ &\le \frac{c_2 M_k e^4}{m^2}\sqrt{\frac{l_k \log(k)}{k}} \cdot \begin{cases} 2 & \text{if } m > m_0, \\ m_0\left(e^2\sqrt{c_2} + 1\right)^{m_0} & \text{if } m \le m_0 \end{cases} \\ &\le \frac{c_6}{m^2}\sqrt{\frac{l_k \log^3(k)}{k}} \end{aligned}$$

for $c_6 = 2c_2e^4(m_0(e^2\sqrt{c_2}+1)^{m_0}+2)$. In the last step, we used that $M_k \le 2\log(k)$ on the event $\mathcal{R}_k$. In a similar fashion, we can establish the bound

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

where $c_7 = 4\sqrt{\lambda}c_3^2(c_3^{2c_3+1}+1)$. Finally, we apply the second claim of Lemma B.7 and obtain

$$\sum_{j=2}^{M_k} \frac{t_j - u_j}{j} \le -C_1' \frac{m-n_k}{n_k m^3} + c_8 \frac{\log(k)}{m^2 k}$$

with $C_1' = (\lambda^2(a+b+1)^2)^{-1}$ and $c_8 = 36(1+\lambda)^3(1+a+b)^2$ for sufficiently large k. We conclude

$$\begin{aligned} f'(m) &\le (T_1 - U_1) + \sum_{j=2}^{M_k} \frac{T_j - t_j}{j} + \sum_{j=2}^{M_k} \frac{u_j - U_j}{j} + \sum_{j=2}^{M_k} \frac{t_j - u_j}{j} \\ &\le -\frac{C_1'}{m^3}\frac{m-n_k}{n_k} + \frac{C_2'}{m^2}\sqrt{\frac{l_k \log^3(k)}{k}} \end{aligned} \tag{5.11}$$

for $C_2' = 4\lambda(a+b) + c_6 + c_7 + c_8$.

Posterior bound. By applying the two inequalities (5.10) and (5.11) for $m < n_k$ and $m > n_k$, we can now bound the posterior probability $\Pi(n \in A_k | \mathbf{X}^k)$ on the event $\mathcal{X}_k$ through equation (5.3). Recall that we can assume $n \ne n_k$ due to the assumption $n_k \notin A_k$. We observe that

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

Noting $|n - n_k| \ge 1$, it also holds that

$$\left|\frac{n}{n_k} - 1\right| = \frac{|n-n_k|}{n_k} \ge \left\{\begin{array}{ll} 1/n_k & \text{if } n_k \le 2n \\ 1/2 & \text{if } n_k > 2n \end{array}\right\} \ge \frac{1}{2n}. \tag{5.12}$$

Therefore, if $l_k \le n < n_k$, the function $h(m)$ introduced in equation (5.10) satisfies

$$\begin{aligned} \int_{n}^{n_k} h(m)\,\mathrm{d}m &= \frac{C_1}{2}\frac{(n-n_k)^2}{(n_k n)^2} - \frac{c_4^5}{2}\frac{n_k^4 - n^4}{(n_k n)^4} \\ &\ge \frac{C_1}{2}\frac{(n-n_k)^2}{(n_k n)^2}\left(1 - \frac{4c_4^5}{C_1}\frac{1}{1-n/n_k}\frac{1}{n^2}\right) \ge 0 \end{aligned} \tag{5.13}$$

for k such that $l_k \ge 8c_4^5/C_1$. Employing bound (5.10) thus yields

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

where the constant C is given by $2C_2/C_1$. On the other hand, for $n_k < n$, bound (5.11) similarly leads to

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

for $C' = 2C_2'/C_1'$. Finally, let $\widetilde{C}_1 = \min\{C_1, C_1'\}$ and $\widetilde{C} = \max\{C, C'\}$. Combining the two inequalities for $n_k < n$ and $n_k > n$ results in

$$k\int_{n_k}^{n} f'(m)\,\mathrm{d}m \le -k\frac{\widetilde{C}_1}{2n_k^2}\Big(\frac{n_k}{n}-1\Big)^2\Big(1-\frac{\widetilde{C}n_k}{|1-n_k/n|}\sqrt{\frac{l_k\log^3(k)}{k}}\Big)$$

for all $n \neq n_k$ with $n \ge M_k$. In order to bound $\Pi(n \in A_k|\mathbf{X}^k)$ via (5.3), we need that the second factor in this expression is positive for large k. Since $l_k\log^3(k) = o(\log^{7/2}(k))$, this motivates the choice

$$A_k := \Big\{n \in \mathbb{N} \big| |n_k - n| \ge n n_k \frac{\log^\rho(k)}{2\sqrt{k}}\Big\} \quad \text{with } \rho = 7/4.$$

For $n \in A_k$ and k large enough, we thus find

$$k\int_{n_k}^{n} f'(m)\,\mathrm{d}m \le -k\frac{\widetilde{C}_1}{4n_k^2}\Big(\frac{n_k}{n}-1\Big)^2$$

Applying the inequalities (5.3) and (5.7) combined with the constraint $\Pi_n(n) \ge \beta n^{-\alpha}$ for all $n \in \mathbb{N}$ on the (proper) prior yields

$$\begin{aligned}\mathbf{1}_{\mathcal{X}_k}\Pi\big(n \in A_k|\mathbf{X}^k\big) &\le 2\lambda c_1 n_k \sum_{n\in A_k}\exp\Big(-\frac{\widetilde{C}_1}{4}\frac{k}{n_k^2}\Big(\frac{n_k}{n}-1\Big)^2\Big)\frac{\Pi_n(n)}{\Pi_n(n_k)}\\ &\le \frac{2\lambda c_1}{\beta}\exp\Big(-\frac{\widetilde{C}_1}{2}\log^{2\rho}(k)+(\alpha+1)\log(n_k)\Big)\longrightarrow 0\end{aligned} \tag{5.14}$$

as $k \to \infty$ uniformly over $\mathcal{M}_k(\lambda)$. Due to (5.8), we have therefore established uniform convergence of $\mathbb{E}_k[\Pi(n \in A_k|\mathbf{X}^k)]$ to 0. To bring this result in the form of Theorem 1, we just have to note that

$$A_k^{\mathrm{c}} \subset \Big\{n \in \mathbb{N}\Big||n_k - n| \le n_k^2\frac{\log^\rho(k)}{\sqrt{k}}\Big\}$$

whenever k is large enough such that $n_k\log^\rho(k)/\sqrt{k} < 1/2$. □

Acknowledgements. We would like to thank the reviewers and are particularly grateful to one referee for a detailed report with additional insights and hints to the literature. These comments have lead to a substantial improvement of the article.

We also thank Alexander Egner and Oskar Laitenberger for providing us with data recorded at the Institute for Nanophotonics Göttingen e.V.

J. Schmidt-Hieber, L. F. Schneider and T. Staudt have contributed equally to this work

Funding. A.M. and T.S. acknowledge support and funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2067/1-390729940, DFG CRC 755 (A6), and DFG RTN 2088. L.F.S. was funded by DFG RTG 2088 (B4) and J.S.H. was supported by a TOP II grant from the NWO.

SUPPLEMENTARY MATERIAL

Supplementary file: Additional proofs and simulations (DOI: 10.1214/21-AOS2096SUPPA; .pdf). Contains the proofs of Theorems 2, 3, and 4 as well as technical lemmas cited in the proof of Theorem 1. Additional figures that depict simulation results akin to the ones in Figures 1 and 4 are also provided.

Supplementary video: Fluorescence microscopy (DOI: 10.1214/21-AOS2096SUPPB; .zip). Video of the first 9000 frames of the data used for estimating the fluorophore number in Section 4 (http://www.stochastik.math.uni-goettingen.de/SMS-movie.mp4).

REFERENCES

ASPELMEIER, T., EGNER, A. and MUNK, A. (2015). Modern statistical challenges in high-resolution fluorescence microscopy. *Annu. Rev. Stat. Appl.* **2** 163–202.

BASU, S. and EBRAHIMI, N. (2001). Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika* **88** 269–279. MR1841274 https://doi.org/10.1093/biomet/88.1.269

BERGER, J. O., BERNARDO, J. M. and SUN, D. (2012). Objective priors for discrete parameter spaces. *J. Amer. Statist. Assoc.* **107** 636–648. MR2980073 https://doi.org/10.1080/01621459.2012.682538

BETZIG, E., PATTERSON, G. H., SOUGRAT, R., LINDWASSER, O. W., OLENYCH, S., BONIFACINO, J. S., DAVIDSON, M. W., LIPPINCOTT-SCHWARTZ, J. and HESS, H. F. (2006). Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313** 1642–1645.

BLUMENTHAL, S. and DAHIYA, R. C. (1981). Estimating the binomial parameter n. *J. Amer. Statist. Assoc.* **76** 903–909. MR0650902

BOCHKINA, N. A. and GREEN, P. J. (2014). The Bernstein-von Mises theorem and nonregular models. *Ann. Statist.* **42** 1850–1878. MR3262470 https://doi.org/10.1214/14-AOS1239

BOUCHERON, S. and GASSIAT, E. (2009). A Bernstein-von Mises theorem for discrete probability distributions. *Electron. J. Stat.* **3** 114–148. MR2471588 https://doi.org/10.1214/08-EJS262

CARROLL, R. J. and LOMBARD, F. (1985). A note on N estimators for the binomial distribution. *J. Amer. Statist. Assoc.* **80** 423–426. MR0792743

CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874 https://doi.org/10.1214/15-AOS1334

CASTILLO, I. and VAN DER VAART, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. MR3059077 https://doi.org/10.1214/12-AOS1029

CHOIRAT, C. and SERI, R. (2012). Estimation in discrete parameter models. *Statist. Sci.* **27** 278–293. MR2963996 https://doi.org/10.1214/11-STS371

DASGUPTA, A. and RUBIN, H. (2005). Estimation of binomial parameters when both n, p are unknown. *J. Statist. Plann. Inference* **130** 391–404. MR2128016 https://doi.org/10.1016/j.jspi.2004.02.019

DEGROOT, M. H. and SCHERVISH, M. J. (2012). *Probability and Statistics*. Pearson Education, Upper Saddle River, NJ.

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

FISHER, R. A. (1941). The negative binomial distribution. *Annu. Eugen.* **11** 182–187. MR0006689

FÖLLING, J., BOSSI, M., BOCK, H., MEDDA, R., WURM, C. A., HEIN, B., JAKOBS, S., EGGELING, C. and HELL, S. W. (2008). Fluorescence nanoscopy by ground-state depletion and single-molecule return. *Nat. Methods* **5** 943–945.

GAO, C., VAN DER VAART, A. W. and ZHOU, H. H. (2020). A general framework for Bayes structured linear models. *Ann. Statist.* **48** 2848–2878. MR4152123 https://doi.org/10.1214/19-AOS1909

GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228

GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782 https://doi.org/10.1017/9781139029834

GÜNEL, E. and CHILKO, D. (1989). Estimation of parameter n of the binomial distribution. *Comm. Statist. Simulation Comput.* **18** 537–551. MR1016224 https://doi.org/10.1080/03610918908812775

HALDANE, J. B. S. (1941). The fitting of binomial distributions. *Annu. Eugen.* **11** 179–181. MR0006688

HALL, P. (1994). On the erratic behavior of estimators of N in the binomial N, p distribution. *J. Amer. Statist. Assoc.* **89** 344–352. MR1266304

HAMEDANI, G. G. and WALTER, G. G. (1988). Bayes estimation of the binomial parameter n. *Comm. Statist. Theory Methods* **17** 1829–1843. MR0945788 https://doi.org/10.1080/03610928808829716

HELL, S. W. (2009). Microscopy and its focal switch. *Nat. Methods* **6** 24–32. https://doi.org/10.1038/nmeth.1291

HELL, S. W. (2015). Nobel lecture: Nanoscopy with freely propagating light. *Rev. Modern Phys.* **87** 1169–1181.

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

HOEL, P. G. (1947). Discriminating between binomial distributions. *Ann. Math. Stat.* **18** 556–564. MR0023035 https://doi.org/10.1214/aoms/1177730346

KAHN, W. D. (1987). A cautionary note for Bayesian estimation of the binomial parameter n. *Amer. Statist.* **41** 38–40. MR0882767 https://doi.org/10.2307/2684316

KARATHANASIS, C., FRICKE, F., HUMMER, G. and HELLEMANN, M. (2017). Molecule counts in localization microscopy with organic fluorophores. *ChemPhysChem* **18** 942–948.

KEMP, A. W. (1997). Characterizations of a discrete normal distribution. *J. Statist. Plann. Inference* **63** 223–229. MR1491581 https://doi.org/10.1016/S0378-3758(97)00020-7

LEE, S.-H., SHIN, J. Y., LEE, A. and BUSTAMANTE, C. (2012). Counting single photoavtiatable fluorescent molecules by photoactivated localization microscopy (palm). *Proc. Natl. Acad. Sci. USA* **109** 17436–17441.

LEHMANN, E. and CASELLA, G. (1996). *Theory of Point Estimation*, 2nd ed. Springer, Berlin.

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

OLKIN, I., PETKAU, A. J. and ZIDEK, J. V. (1980). A comparison of *n* estimators for the binomial distribution. *J. Amer. Statist. Assoc.* **76** 637–642.

OTIS, D. L., BURNHAM, K. P., WHITE, G. C. and ANDERSON, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildl. Monogr.* **64** 1–135.

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

ROLLINS, G. C., SHIN, J. Y., BUSTAMANTE, C. and PRESSÉ, S. (2015). Stochastic approach to the molecular counting problem in superresolution microscopy. *Proc. Natl. Acad. Sci. USA* **112** E110–8.

ROYLE, J. A. (2004). *N*-mixture models for estimating population size from spatially replicated counts. *Biometrics* **60** 108–115. MR2043625 https://doi.org/10.1111/j.0006-341X.2004.00142.x

RUST, M. J., BATES, M. and ZHUANG, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3** 793–796.

SCHMIDT-HIEBER, J., SCHNEIDER, L. F., STAUDT, T., KRAJINA, A., ASPELMEIER, T. and MUNK, A. (2021). Supplement to "Posterior analysis of *n* in the binomial (*n*, *p*) problem with both parameters unknown—with applications to quantitative nanoscopy." https://doi.org/10.1214/21-AOS2096SUPPA, https://doi.org/10.1214/21-AOS2096SUPPB

SCHMIED, J. J., RAAB, M., FORTHMANN, C., PIBIRI, E., WÜNSCH, B., DAMMEYER, T. and TINNEFELD, P. (2014). Dna origami-based standards for quantitative fluoresence microscopy. *Nat. Protoc.* **9** 1367–1391.

SCHNEIDER, L. F., STAUDT, T. and MUNK, A. (2018). Posterior consistency in the binomial model with unknown parameters: A numerical study. In *International Conference on Bayesian Statistics in Action* 35–42. Springer, Berlin.

SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26. MR0184378 https://doi.org/10.1007/BF00535479

SMITH, P. J. (1988). Bayesian methods for multiple capture-recapture surveys. *Biometrics* **44** 1177–1189.

STAUDT, T., ASPELMEIER, T., LAITENBERGER, O., GEISLER, C., EGNER, A. and MUNK, A. (2020). Statistical molecule counting in super-resolution fluorescence microscopy: Towards quantitative nanoscopy. *Statist. Sci.* **35** 92–111. MR4071360 https://doi.org/10.1214/19-STS753

STUDENT (1919). An explanation of deviations from Poisson's law in practice. *Biometrika* **12** 211–215.

SZABŁOWSKI, P. (2001). Discrete normal distribution and its relationship with Jacobi theta functions. *Statist. Probab. Lett.* **52** 289–299. MR1838217 https://doi.org/10.1016/S0167-7152(00)00223-6

TA, H., KELLER, J., HALTMEIER, M., SAKA, S. K., SCHMIED, J., OPAZO, F., TINNEFELD, P., MUNK, A. and HELL, S. W. (2015). Mapping molecules in scanning far-field fluorescence nanoscopy. *Nat. Commun.* **6** 1–7.

TANCREDI, A., STEORTS, R. and LISEO, B. (2020). A unified framework for de-duplication and population size estimation (with discussion). *Bayesian Anal.* **15** 633–682. MR4122517 https://doi.org/10.1214/19-BA1146

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

VILLA, C. and WALKER, S. G. (2014). A cautionary note on the discrete uniform prior for the binomial n: Comment. *Ecology* **95** 2674–2677.

VILLA, C. and WALKER, S. G. (2015). An objective approach to prior mass functions for discrete parameter spaces. *J. Amer. Statist. Assoc.* **110** 1072–1082. MR3420685 https://doi.org/10.1080/01621459.2014.946319

WANG, X., HE, C. Z. and SUN, D. (2007). Bayesian population estimation for small sample capture-recapture data using noninformative priors. *J. Statist. Plann. Inference* **137** 1099–1118. MR2301466 https://doi.org/10.1016/j.jspi.2006.03.004