International Journal of Advanced Statistics and Probability, 13 (2) (2026) 145-152

International Journal of Advanced Statistics and Probability

Website: www.sciencepubco.com/index.php/IJASP
https://doi.org/10.14419/qf3g7k08
Research paper

Frequentist Versus Bayesian Parameter Estimation of The Beta Distribution: A Simulation-Based Analysis

Felix James [1] *, Ukaogo Greatman Odinakachukwu [1], Ohanuba Felix Obi [2]

[1] *Department of Data Science and Analytics, School of Information Technology and Computing, American University of Nigeria, Yola, Adamawa State*
[2] *Department of Statistics, University of Nigeria, Nsukka, Nigeria*
**Corresponding author. E-mail: felix.james@aun.edu.ng*

Received: 26 Aug 2026. Accepted: 16 Oct 2026. Published: 2 Nov 2026

Abstract

The Beta distribution is a highly useful continuous probability distribution defined on the interval (0, 1) and characterized by two shape parameters, α and β. Accurate estimation of these parameters is crucial in various fields, ranging from reliability engineering to Bayesian A/B testing and click-through rate smoothing in data science. This paper presents a comprehensive comparative analysis of three parameter estimation frameworks: the Method of Moments (MOM), Maximum Likelihood Estimation (MLE), and Bayesian estimation via Maximum A posteriori (MAP) under a Gamma prior. Building on previous investigations which demonstrated the superiority of MOM in small sample sizes (N<50), this study conducts extensive Monte Carlo simulations across multiple scenarios (symmetric, skewed, and U-shaped distributions) and sample sizes ranging from 10 to 500. Our results confirm that MOM remains computationally trivial and has lower mean squared error (MSE) than MLE in extremely small samples (N=10), while MLE is asymptotically efficient and excels in large samples. The Bayesian MAP estimator with a Gamma(2, 0.5) prior outperforms MLE in small-to-moderate samples in the skewed and symmetric scenarios, with MSE reductions of roughly 90% in the most favourable case, and converges to MLE as the sample size increases. A sensitivity analysis over six Gamma priors shows, however, that this gain depends on how well the prior is located relative to the true parameters: poorly located priors give little improvement or are worse than MLE, most clearly in the U-shaped scenario. Convergence diagnostics over 21,000 simulated datasets show no optimizer failures or boundary solutions for MLE or MAP and no variance violations for MOM. We discuss these trade-offs and provide concrete recommendations for practitioners.

Keywords: *Beta Distribution; Parameter Estimation; Method of Moments; Maximum Likelihood Estimation; Bayesian MAP; Monte Carlo Simulation; Prior Sensitivity.*

1. Introduction

The Beta distribution is a family of continuous probability distributions defined on the interval (0, 1) and parametrised by two positive shape parameters, α and β. Owing to its ability to take on a wide variety of shapes, including symmetric, highly skewed, uniform, and U-shaped density curves, the Beta distribution is widely used to model continuous random variables restricted to a finite interval [2]. Typical applications include modeling percentages, proportions, and physical limits such as project completion times in PERT networks, soil moisture contents in hydrology, and reliability indices in engineering.

Despite its utility, estimating the parameters of the Beta distribution can be numerically challenging. Because the probability density function involves beta functions and gamma terms, direct closed-form estimation is often impossible under classical frameworks. In our previous work, Investigating the Parameters of the Beta Distribution [1], we comprehensively reviewed classical estimation techniques. We formulated the system of likelihood equations for Maximum Likelihood Estimation (MLE) and contrasted them with the analytical Method of Moments (MOM) estimators. A key discovery of that investigation was that MOM estimators are highly preferable for small sample sizes (specifically samples under 50) because they are straightforward, computationally instantaneous, and avoid the numerical instability or non-convergence that plagues the numerical maximization of the likelihood function. Conversely, MLE was shown to be superior in large samples due to its asymptotic consistency and efficiency, achieving the Cramér-Rao lower bound on variance.

Following our 2023 study, we observed that while MOM is highly practical for manual calculation in small samples, it lacks the flexibility to incorporate prior domain knowledge. In fields like digital advertising and clinical trials, practitioners often possess historical baseline data or expert opinions. The need to bridge that gap directly motivated this paper using a Bayesian framework. While classical closed-form alternatives have been recently investigated [4 - 7], they often lack the capacity to easily incorporate prior regularizing beliefs directly into the optimization process.

In modern data science and statistical modeling, however, there is a growing need to handle small-sample regimes without sacrificing the statistical properties that make MLE attractive. This is particularly relevant in online web applications, such as click-through rate (CTR) estimation, Thompson Sampling for multi-armed bandit systems, and Bayesian A/B testing conducted by technology firms [3] (e.g., Etsy,

Copyright © Felix James et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Google, and Amazon). In these contexts, sparsity of data is a constant challenge; new articles, advertisements, or product listings have very few impressions (low sample sizes), leading to highly noisy and unstable point estimates if classical MOM or MLE are used directly.
To address this limitation, this paper expands the scope of [1] by integrating the Bayesian paradigm. Bayesian estimation allows for the incorporation of prior beliefs about the parameters via prior distributions, which stabilizes estimation in sparse data settings by shrinking noisy point estimates toward a global or historical mean. Specifically, we evaluate the Maximum A posteriori (MAP) estimator under weekly informative Gamma priors on the shape parameters. Through an extensive Monte Carlo simulation study, we evaluate the performance of MOM, MLE, and Bayesian MAP across three shape scenarios (symmetric, highly skewed, and U-shaped) and sample sizes ranging from 10 to 500. Because the performance of a MAP estimator necessarily depends on the prior, we also report a sensitivity analysis over six Gamma priors and full convergence diagnostics for the numerical optimizers. This study provides statisticians and data scientists with an evidence-based guide to choosing an estimator according to sample size, underlying population shape, and the quality of the available prior information.

2. Mathematical Formulation of Estimators

To establish a foundation for the comparative simulation, we present the mathematical formulations for the three estimation methodologies. Let $X = \{x_1, x_2, \dots, x_N\}$ be an independent and identically distributed (i. i. d.) random sample of size N drawn from a Beta distribution, denoted as $X \sim \text{Beta}(\alpha, \beta)$.
The probability density function is given by:

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \qquad 0 < x < 1$$

Where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the Beta function, and $\Gamma(\bullet)$ Is the Gamma function.

2.1. Method of moments (MOM) estimators.

The Method of Moments matches the theoretical moments of the Beta distribution with their sample counterparts. The sample mean and sample variance are defined respectively as

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{and} \qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

By equating these to the theoretical mean and variance of the Beta distribution, we obtain the closed-form MOM estimators:

$$\hat{\alpha}_{MOM} = \bar{x}\left[\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right]$$

$$\hat{\beta}_{MOM} = (1-\bar{x})\left[\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right]$$

MOM estimators are computationally trivial to evaluate and require no iterative numerical techniques. However, they suffer from two major theoretical drawbacks. First, they are not asymptotically efficient. Second, in extremely small samples, it is mathematically possible for the sample variance to exceed the theoretical maximum variance $\bar{x}(1 - \bar{x})$, which leads to invalid negative parameter estimates. In practice, these must be truncated to a small positive boundary value (e.g., 10^{-5}).

2.2 Maximum likelihood estimation (MLE)

Maximum Likelihood Estimation seeks to find parameter values that maximize the likelihood of the observed data. The joint log-likelihood function for the sample is formulated as:

$$\ln L(\alpha, \beta; X) = N\ln\Gamma(\alpha+\beta) - N\ln\Gamma(\alpha) - N\ln\Gamma(\beta) + (\alpha-1)\sum_{i=1}^{N}\ln x_i + (\beta-1)\sum_{i=1}^{N}\ln(1-x_i)$$

Taking the partial derivatives of the log-likelihood with respect to α and β gives the score functions.

$$\frac{\partial \ln L}{\partial \alpha} = N[\psi(\alpha+\beta) - \psi(\alpha)] + \sum_{i=1}^{N}\ln x_i, \qquad \frac{\partial \ln L}{\partial \beta} = N[\psi(\alpha+\beta) - \psi(\beta)] + \sum_{i=1}^{N}\ln(1-x_i),$$

Where $\psi(z) = \frac{d}{dz}\ln\Gamma(z) = \Gamma'(z)/\Gamma(z)$ is the digamma function. Setting both score functions to zero yields the system of likelihood equations:

$$\psi(\hat{\alpha}) - \psi(\hat{\alpha}+\hat{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\ln x_i, \qquad \psi(\hat{\beta}) - \psi(\hat{\alpha}+\hat{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\ln(1-x_i).$$

Because this system is non-linear and coupled, it has no closed-form solution, and numerical optimization (for example, L-BFGS-B or Newton-Raphson) is required. Newton-Raphson uses the Hessian of the log-likelihood,

$$H(\alpha, \beta) = N\begin{pmatrix} \psi_1(\alpha+\beta) - \psi_1(\alpha) & \psi_1(\alpha+\beta) \\ \psi_1(\alpha+\beta) & \psi_1(\alpha+\beta) - \psi_1(\beta) \end{pmatrix},$$

Where $\psi_1(z) = \psi'(z)$ is the trigamma function. H is negative definite for all $\alpha, \beta > 0$, so the log-likelihood is strictly concave and has at most one maximizer. As highlighted in [1], numerical optimization can nevertheless be unstable in small samples, in which case the maximizer may lie at very large parameter values; we therefore report convergence diagnostics explicitly (Sections 3.1 and 4.5).

2.3. Bayesian estimation and maximum a posteriori (MAP)

Under the Bayesian paradigm, the parameters are treated as random variables. We specify independent prior distributions for α and β. Since shape parameters must be strictly positive, we assign them Gamma prior distributions:

$\alpha \sim \mathrm{Gamma}(a_1, b_1)$ and $\beta \sim \mathrm{Gamma}(a_2, b_2)$

Where the probability density of a Gamma-distributed variable is proportional to $z^{a-1}\exp(-bz)$. According to Bayes' theorem, the joint posterior distribution is proportional to the product of the likelihood and the priors:

$p(\alpha, \beta \mid X) \propto L(\alpha, \beta; X)\, p(\alpha; a_1, b_1)\, p(\beta; a_2, b_2)$

By taking the logarithm, we obtain the log-posterior distribution (excluding the normalizing constant):

$\ln p(\alpha, \beta \mid X) = \ln L(\alpha, \beta; X) + (a_1 - 1)\ln\alpha - b_1\alpha + (a_2 - 1)\ln\beta - b_2\beta$

We selected independent Gamma(2, 0.5) priors for both α and β, where Gamma(a, b) denotes the shape-rate parameterization used above. This prior has mode 2.0, mean 4.0 and standard deviation ≈ 2.83, and places about 9% of its mass below 1. It acts as a regularizer that prevents the numerical optimizer from drifting into extreme, unstable parameter values when the sample is small, while allowing the likelihood to dominate as N increases. It is weakly informative for moderate shape values such as those in the skewed and symmetric scenarios, but it is not neutral when the true shape parameters are below 1 (the U-shaped scenario). Section 4.4 therefore examines five further Gamma priors. All priors considered have shape $a \geq 1$, so the log-prior is concave and, together with the strictly concave log-likelihood, the log-posterior has a unique maximum. Its gradient is the score above plus $(a_1 - 1)/\alpha - b_1$ and $(a_2 - 1)/\beta - b_2$ For the α and β components, respectively.

The Maximum A Posteriori (MAP) estimator is the mode of the joint posterior distribution, obtained by numerically maximizing the log-posterior. The addition of the Gamma prior terms acts as a mathematical penalty or regularizer. When the sample size is small, the prior terms prevent the estimates from taking extreme values, substantially reducing the variance of the estimator. As the sample size N grows, the likelihood term dominates the prior terms, and the Bayesian MAP estimator converges to the classical MLE.

3. Monte Carlo simulation methodology

To rigorously evaluate the comparative performance of MOM, MLE, and Bayesian MAP estimators, we designed and executed a large-scale Monte Carlo simulation in Python. The simulation was structured around three key shape scenarios representing diverse distributional properties:

1) Skewed Scenario ($\alpha = 2.0, \beta = 5.0$): Represents a highly right-skewed distribution, typical of CTR and conversion rate data in web systems.
2) Symmetric Scenario ($\alpha = 2.0, \beta = 2.0$): Represents a symmetric, unimodal bell-shaped distribution centered at 0.5.
3) U-Shaped Scenario ($\alpha = 0.5, \beta = 0.5$): Represents a U-shaped distribution where data is heavily concentrated near the boundaries (0 and 1).

For each scenario, we simulated datasets of varying sample sizes, capturing both small and large-sample behaviors: $N \in \{10, 20, 30, 50, 100, 250, 500\}$. Each experimental condition was replicated $R = 1{,}000$ times to ensure stable statistical metrics. For the Bayesian MAP estimator, we assigned weakly informative Gamma priors on both shape parameters with parameters $a_1 = a_2 = 2.0$ and $b_1 = b_2 = 0.5$. This Gamma prior has a mode of 2.0 and a mean of 4.0, which reflects a mild, positive-support regularizer centered in the typical range of shape parameters without being overly restrictive.

The performance of each estimator was rigorously evaluated using two key metrics: Bias and Mean Squared Error (MSE). These are defined respectively as follows (using θ to represent either parameter):

3.1. Computational implementation and convergence diagnostics.

All estimators were implemented in Python using NumPy and SciPy. MOM was computed in closed form; whenever the sample variance violated $s^2 < \bar{x}(1 - \bar{x})$ (so that the bracketed term in Section 2.1 was non-positive), the estimate was truncated at 10^{-5} and the replication was recorded as a MOM variance violation. MLE and MAP were obtained by minimizing the negative log-likelihood and the negative log-posterior, respectively, with the L-BFGS-B algorithm (copy.optimize.minimize), using the analytic gradients of Sections 2.2 and 2.3, box constraints $10^{-5} \leq \alpha, \beta \leq 1000$ and at most 500 iterations. Every optimization was started from two initial values, the (truncated) MOM estimate and $(\alpha, \beta) = (1,1)$, and the solution with the smaller objective value was retained. Because the objective has a unique optimum (Sections 2.2 and 2.3), the second start serves as a check on the optimizer rather than as a search for a global optimum.

For every fit we recorded (i) non-convergence, when the optimizer reported failure for the retained solution; (ii) a boundary solution, when either estimate was within 10^{-4} of the lower bound or within 1% of the upper bound; and (iii) start disagreement, when fewer than both starting values converged. Within each scenario, sample size and replication, the same simulated dataset was supplied to every estimator and every prior, so all comparisons are paired, and random seeds were fixed for reproducibility. The results are summarized in Section 4.5.

3.2. Design of the prior sensitivity analysis.

The baseline analysis uses a single Gamma(2, 0.5) prior for both shape parameters. To assess robustness to this choice, the whole simulation was repeated, on the same datasets, with the five additional priors listed in Table 4, applied to both parameters. They include a vague prior (Gamma(1, 0.1)), priors centered at or below the true values in the skewed and symmetric scenarios (Gamma(1, 1), Gamma(2, 1)), and priors centered well above them (Gamma(2, 0.1), Gamma(5, 1)); Gamma(1, 1) is also the prior closest to the U-shaped truth. Performance is summarized for each prior by the mean of MSE($\hat{\alpha}$) and MSE($\hat{\beta}$).

$\text{Bias}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta)$ and $\text{MSE}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta)^2$

4. Simulation Results

The empirical findings of our Monte Carlo simulation are informative. Below, we present the quantitative results for the three distributional scenarios. Let us first examine the right-skewed scenario, which represents a highly common distributional shape in industrial data science applications.

4.1. Right-Skewed Scenario (α = 2.0, β = 5.0)

Table 1 summarizes the mean estimates, bias, and MSE of the three estimation methodologies across all sample sizes for the skewed distribution.

Table 1: Mean Estimates, Bias, and MSE For the Right-Skewed Scenario (A = 2.0, B = 5.0). Bayes-MAP Uses the Gamma(2, 0.5) Prior

N	Method	Mean α (true=2.0)	Mean β (true=5.0)	Bias α	Bias β	MSE α	MSE β
10	Bayes-MAP	1.965	4.667	-0.035	-0.333	0.294	1.487
10	MLE	2.810	7.174	0.810	2.174	3.644	23.043
10	MOM	2.517	6.431	0.517	1.431	2.811	18.296
20	Bayes-MAP	1.979	4.863	-0.021	-0.137	0.211	1.345
20	MLE	2.295	5.852	0.295	0.852	0.622	4.682
20	MOM	2.180	5.548	0.180	0.548	0.598	4.433
30	Bayes-MAP	2.000	4.947	0.000	-0.053	0.185	1.194
30	MLE	2.201	5.575	0.201	0.575	0.394	2.837
30	MOM	2.122	5.368	0.122	0.368	0.399	2.818
50	Bayes-MAP	2.010	4.983	0.010	-0.017	0.123	0.863
50	MLE	2.121	5.334	0.121	0.334	0.192	1.416
50	MOM	2.072	5.207	0.072	0.207	0.196	1.384
100	Bayes-MAP	2.004	4.995	0.004	-0.005	0.067	0.448
100	MLE	2.055	5.157	0.055	0.157	0.082	0.569
100	MOM	2.042	5.120	0.042	0.120	0.095	0.630
250	Bayes-MAP	2.005	4.996	0.005	-0.004	0.028	0.194
250	MLE	2.024	5.058	0.024	0.058	0.030	0.212
250	MOM	2.017	5.038	0.017	0.038	0.035	0.234
500	Bayes-MAP	2.000	4.997	-0.000	-0.003	0.015	0.104
500	MLE	2.009	5.028	0.009	0.028	0.015	0.108
500	MOM	2.005	5.017	0.005	0.017	0.018	0.122

The quantitative findings in Table 1 reveal an extraordinary performance advantage for the Bayesian MAP estimator in small samples. For N = 10, the classical Maximum Likelihood Estimator exhibits massive instability, resulting in a Mean Squared Error of 3.644 for α and a staggering 23.043 for β. While the Method of Moments is slightly more stable than MLE, its MSE is still very high (2.811 for α and 18.296 for β). In contrast, the Bayesian MAP estimator achieves an MSE of only 0.294 for α and 1.487 for β, a reduction in MSE of roughly 90-94% relative to the two classical frameworks. This demonstrates the profound stabilizing impact of weakly informative priors in the sparse-data regime.

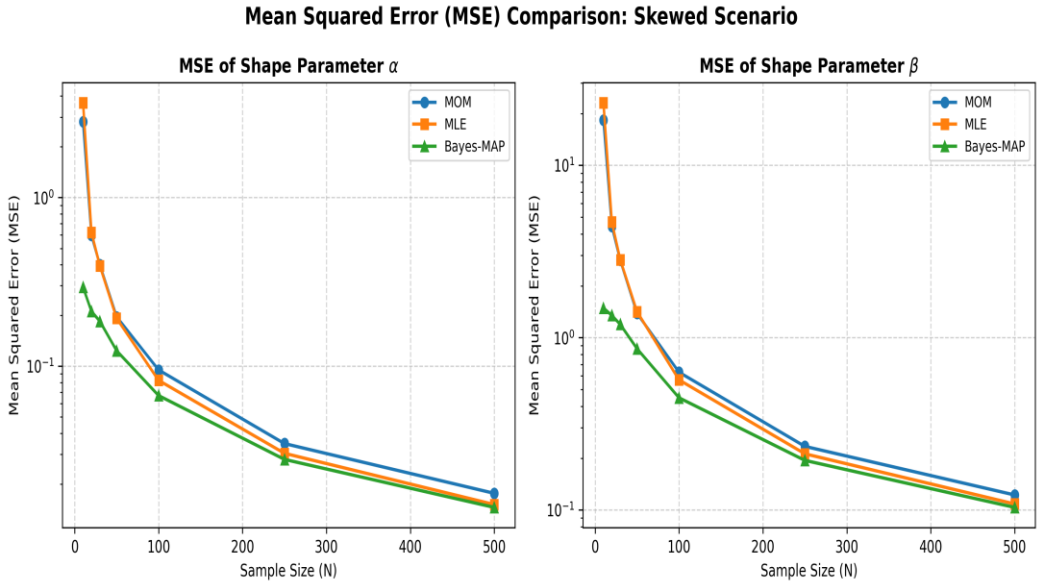


Fig. 1: Mean Squared Error (MSE) Comparison for Skewed Scenario (α = 2.0, β = 5.0).

4.2. Symmetric scenario (α = 2.0, β = 2.0)

Table 2 presents the statistical metrics for the symmetric bell-shaped distribution across sample sizes.

Table 2: Mean Estimates, Bias and MSE For the Symmetric Scenario (A = 2.0, B = 2.0). Bayes-MAP Uses the Gamma(2, 0.5) Prior

N	Method	Mean α (true=2.0)	Mean β (true=2.0)	Bias α	Bias β	MSE α	MSE β
10	Bayes-MAP	2.372	2.395	0.372	0.395	0.794	0.837
10	MLE	2.764	2.791	0.764	0.791	3.304	3.272
10	MOM	2.365	2.391	0.365	0.391	2.310	2.252
20	Bayes-MAP	2.251	2.247	0.251	0.247	0.427	0.443
20	MLE	2.339	2.336	0.339	0.336	0.701	0.736
20	MOM	2.162	2.160	0.162	0.160	0.588	0.615
30	Bayes-MAP	2.180	2.182	0.180	0.182	0.290	0.313
30	MLE	2.220	2.222	0.220	0.222	0.398	0.424
30	MOM	2.108	2.111	0.108	0.111	0.343	0.371
50	Bayes-MAP	2.129	2.123	0.129	0.123	0.172	0.171
50	MLE	2.144	2.137	0.144	0.137	0.204	0.202
50	MOM	2.068	2.062	0.068	0.062	0.193	0.194
100	Bayes-MAP	2.062	2.061	0.062	0.061	0.084	0.085
100	MLE	2.065	2.065	0.065	0.065	0.090	0.091
100	MOM	2.029	2.030	0.029	0.030	0.091	0.094
250	Bayes-MAP	2.019	2.016	0.019	0.016	0.028	0.031
250	MLE	2.020	2.017	0.020	0.017	0.029	0.031
250	MOM	2.009	2.005	0.009	0.005	0.032	0.034
500	Bayes-MAP	2.009	2.009	0.009	0.009	0.014	0.015
500	MLE	2.009	2.009	0.009	0.009	0.014	0.015
500	MOM	2.002	2.002	0.002	0.002	0.015	0.016

In the symmetric scenario (Table 2), we see a similar stabilizing trend. At N = 10, the classical MLE has high MSE (3.304 and 3.272), whereas Bayesian MAP controls the error variance, producing a significantly lower MSE (0.794 and 0.837). MOM is reasonably stable here but exhibits a larger variance than the MAP estimator. As N grows, the results of all three methods merge.

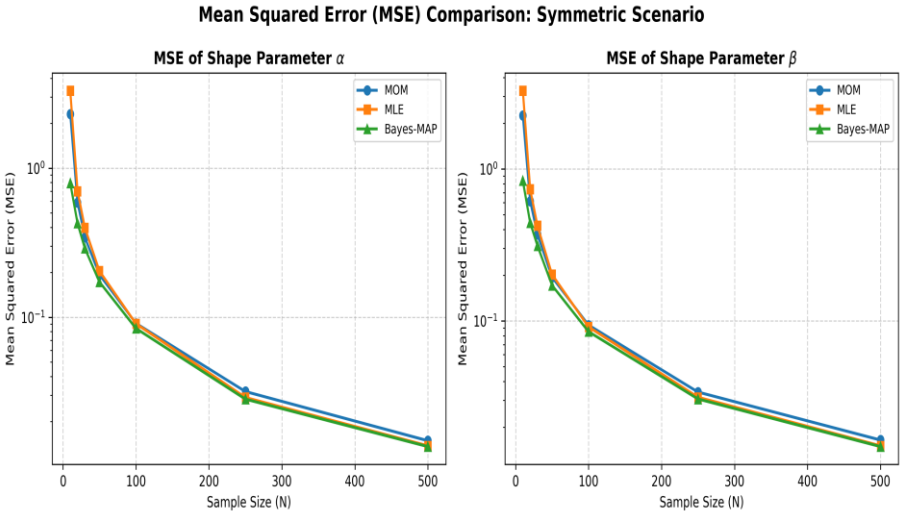


Fig. 2: Mean Squared Error (MSE) Comparison for Symmetric Scenario (α = 2.0, β = 2.0).

4.3. U-Shaped Scenario (α = 0.5, β = 0.5)

Table 3 summarizes the performance metrics for the U-shaped distribution scenario.

Table 3: Mean Estimates, Bias and MSE for the U-Shaped Scenario (A = 0.5, B = 0.5). Bayes-MAP Uses the Gamma(2, 0.5) Prior

N	Method	Mean α (true=0.5)	Mean β (true=0.5)	Bias α	Bias β	MSE α	MSE β
10	Bayes-MAP	0.774	0.763	0.274	0.263	0.243	0.224
10	MLE	0.683	0.672	0.183	0.172	0.217	0.200
10	MOM	0.542	0.534	0.042	0.034	0.175	0.185
20	Bayes-MAP	0.613	0.622	0.113	0.122	0.050	0.055
20	MLE	0.566	0.573	0.066	0.073	0.037	0.041
20	MOM	0.504	0.511	0.004	0.011	0.040	0.043
30	Bayes-MAP	0.562	0.568	0.062	0.068	0.025	0.025
30	MLE	0.531	0.537	0.031	0.037	0.020	0.020
30	MOM	0.493	0.496	-0.007	-0.004	0.027	0.024
50	Bayes-MAP	0.544	0.540	0.044	0.040	0.013	0.012
50	MLE	0.525	0.521	0.025	0.021	0.011	0.010
50	MOM	0.502	0.498	0.002	-0.002	0.014	0.013
100	Bayes-MAP	0.523	0.522	0.023	0.022	0.005	0.005
100	MLE	0.514	0.513	0.014	0.013	0.005	0.005
100	MOM	0.501	0.500	0.001	0.000	0.006	0.006
250	Bayes-MAP	0.507	0.507	0.007	0.007	0.002	0.002
250	MLE	0.503	0.504	0.003	0.004	0.002	0.002
250	MOM	0.500	0.500	-0.000	0.000	0.003	0.003
500	Bayes-MAP	0.507	0.505	0.007	0.005	0.001	0.001
500	MLE	0.505	0.504	0.005	0.004	0.001	0.001
500	MOM	0.503	0.501	0.003	0.001	0.001	0.001

In the U-shaped scenario (Table 3), where the true shape values are small ($\alpha = \beta = 0.5$), the Gamma(2, 0.5) prior, whose mode is 2.0, pulls the estimates upward and introduces a positive bias in small samples (mean bias of about +0.27 at $N = 10$, against +0.18 for MLE and +0.04 for MOM). Consequently, MOM and MLE have lower MSE than Bayes-MAP at $N = 10$, and MLE also has lower MSE than Bayes-MAP at $N = 20$, 30 and 50 (for example 0.037 versus 0.050 for α at $N = 20$). The differences become negligible only from $N = 100$. In this scenario,, the prior is therefore a liability rather than an asset in small samples, and the favorable behavior seen in Tables 1 and 2 should not be extrapolated to it; this is examined further in Section 4.4.

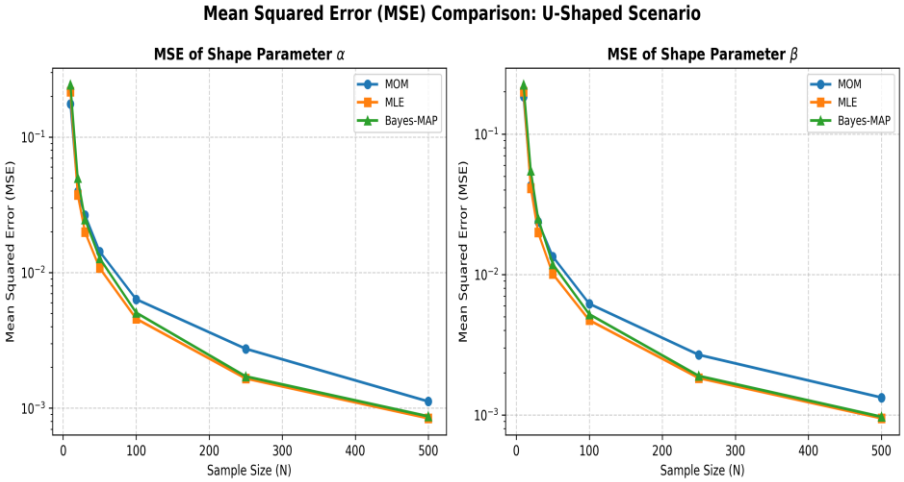


Fig. 3: Mean Squared Error (MSE) Comparison for U-Shaped Scenario ($\alpha = 0.5$, $\beta = 0.5$).

4.4. Prior sensitivity analysis

Table 4 lists the six priors and Table 5 reports the mean MSE of the three estimators under each of them. All comparisons are paired: the same datasets are used for every row of a given scenario and sample size. The Gamma(2, 0.5) rows of Table 5 reproduce Tables 1-3 up to Monte Carlo error, since they come from an independent set of random datasets.

Table 4: Gamma Priors Used in the Sensitivity Analysis (Shape-Rate Parameterization, Applied to Both A and B).

Prior (shape, rate)	Mode	Mean	SD	Prior mass below 1	Role
Gamma(2, 0.5)	2.0	4.0	2.83	9.0%	Baseline (Sections 2-4.3)
Gamma(1, 0.1)	0.0	10.0	10.00	9.5%	Vague (exponential)
Gamma(1, 1)	0.0	1.0	1.00	63.2%	Exponential, centred low
Gamma(2, 1)	1.0	2.0	1.41	26.4%	Moderate, centred low
Gamma(2, 0.1)	10.0	20.0	14.14	0.5%	Diffuse, centred high
Gamma(5, 1)	4.0	5.0	2.24	0.4%	Informative, centred high

Table 5: Mean MSE (Average of MSE($\hat{\alpha}$) and MSE($\hat{\beta}$)) of MLE, MOM and Bayes-MAP Under Six Gamma Priors.

Estimator	N = 10	N = 20	N = 30	N = 50	N = 100	N = 250	N = 500
Right-skewed ($\alpha = 2.0$, $\beta = 5.0$)							
MLE	17.56	3.215	1.566	0.784	0.356	0.120	0.056
MOM	13.15	2.845	1.515	0.801	0.366	0.135	0.062
MAP, Gamma(2, 0.5)	0.964	0.882	0.666	0.492	0.276	0.109	0.054
MAP, Gamma(1, 0.1)	4.603	1.998	1.154	0.665	0.325	0.116	0.055
MAP, Gamma(1, 1)	4.085	2.003	1.223	0.708	0.313	0.116	0.057
MAP, Gamma(2, 1)	2.128	1.144	0.752	0.505	0.261	0.107	0.054
MAP, Gamma(2, 0.1)	13.29	3.871	1.918	0.914	0.393	0.126	0.057
MAP, Gamma(5, 1)	0.992	0.892	0.694	0.506	0.289	0.112	0.054
Symmetric ($\alpha = 2.0$, $\beta = 2.0$)							
MLE	3.717	0.675	0.395	0.202	0.076	0.030	0.016
MOM	2.641	0.571	0.343	0.192	0.083	0.032	0.017
MAP, Gamma(2, 0.5)	0.845	0.409	0.294	0.172	0.071	0.029	0.016
MAP, Gamma(1, 0.1)	1.826	0.529	0.336	0.184	0.073	0.029	0.016
MAP, Gamma(1, 1)	0.507	0.278	0.180	0.124	0.062	0.027	0.015
MAP, Gamma(2, 1)	0.267	0.210	0.167	0.121	0.060	0.027	0.015
MAP, Gamma(2, 0.1)	4.874	1.002	0.560	0.258	0.088	0.032	0.016
MAP, Gamma(5, 1)	2.138	0.907	0.605	0.300	0.104	0.035	0.017
U-shaped ($\alpha = 0.5$, $\beta = 0.5$)							
MLE	0.178	0.043	0.024	0.012	0.0050	0.0017	0.0009
MOM	0.130	0.043	0.026	0.015	0.0066	0.0025	0.0013
MAP, Gamma(2, 0.5)	0.211	0.056	0.029	0.015	0.0055	0.0018	0.0009
MAP, Gamma(1, 0.1)	0.152	0.040	0.023	0.012	0.0049	0.0017	0.0009
MAP, Gamma(1, 1)	0.053	0.027	0.018	0.010	0.0046	0.0016	0.0008
MAP, Gamma(2, 1)	0.126	0.044	0.025	0.013	0.0052	0.0017	0.0009
MAP, Gamma(2, 0.1)	0.345	0.069	0.033	0.016	0.0057	0.0018	0.0009
MAP, Gamma(5, 1)	0.707	0.166	0.073	0.031	0.0090	0.0023	0.0010

Shaded cells: MAP has a larger mean MSE than MLE. The highlighted rows are the baseline Gamma(2, 0.5) prior. Each cell is based on 1,000 replications.

Skewed scenario. At $N = 10$ five of the six priors reduce the mean MSE relative to MLE by between 74% (Gamma(1, 0.1)) and 95% (Gamma(2, 0.5)). Only the diffuse, upward-centred Gamma(2, 0.1) gives a small gain (24%), and by $N = 20$ it is 20% worse than MLE. The gains of the better-located priors persist at $N = 50$ (Gamma(2, 0.5), Gamma(2, 1) and Gamma(5, 1) reduce MSE by 36% to 37%) and are small by $N = 250$.

Symmetric scenario. Gamma(2, 1), Gamma(1, 1) and Gamma(2, 0.5) reduce MSE at $N = 10$ by 93%, 86% and 77%, respectively. The diffuse, upward-centred Gamma(2, 0.1) is 31% worse than MLE at $N = 10$, and at $N = 20$ both Gamma(2, 0.1) and the informative Gamma(5, 1) are worse than MLE (49% and 34% higher MSE).

U-shaped scenario. Here the baseline prior is a poor choice: Gamma(2, 0.5) has 18% higher mean MSE than MLE at $N = 10$ and 32% higher at $N = 20$, and the informative Gamma(5, 1) prior has 3.9 times the MSE of MLE at $N = 20$. Only Gamma(1, 1), the prior with most mass near the true value 0.5, improves on MLE at every sample size up to $N = 100$, reducing MSE by 70% at $N = 10$. The positive small-sample bias of the baseline prior is also visible: at $N = 10$ the mean bias of the baseline MAP estimator is +0.25, against +0.16 for MLE.

By $N = 500$ the MSE of MLE and of all priors except Gamma(5, 1) agree to within about 7%, and the differences between estimators have narrowed considerably by $N = 100$; the strongly informative Gamma(5, 1) prior remains up to about 17% worse than MLE at $N = 500$ in the symmetric and U-shaped scenarios. The size of the MAP advantage is therefore governed by how well the prior is located relative to the true parameters and by how small the sample is: when the prior has substantial mass near the truth the gain can be large, but a prior centred in the wrong place can make MAP worse than MLE or MOM. The roughly 90% MSE reduction reported for the skewed scenario in Table 1 should accordingly be read as the outcome for a prior whose mean (4.0) lies between the true values $\alpha = 2$ and $\beta = 5$, and not as a general property of MAP estimation with weakly informative priors.

4.5. Convergence and boundary diagnostics

Table 6 reports the diagnostics defined in Section 3.1, aggregated over the seven sample sizes. Across the 21,000 simulated datasets, L-BFGS-B reported convergence for every one of the 147,000 MLE and MAP fits (one MLE fit and six MAP fits per dataset); no estimate lay on a bound, and both starting values converged in every case. No MOM variance violation (and hence no truncation at 10^{-5}) occurred in any dataset, although such violations are possible in principle in very small samples. In our implementation, therefore, the poor small-sample performance of MLE and MOM seen in Tables 1 and 2 is statistical rather than numerical: it arises from very large and highly variable estimates (for instance, the mean MLE of β at $N = 10$ in Table 1 is 7.17 against a true value of 5.0), not from failed optimizations or boundary truncation. This conclusion is specific to the implementation used (analytic gradients, box-constrained L-BFGS-B and a moment-based starting value); optimizers without gradients, without bounds or with poor starting values may fail more often in small samples, as noted in [1].

Table 6: Numerical Diagnostics, Aggregated Over N = 10, 20, 30, 50, 100, 250, 500 (1,000 Replications Each). MLE: 7,000 Fits Per Scenario; MAP: Six Priors, 42,000 Fits Per Scenario

Scenario	Datasets	MOM variance violations	MLE non-convergence	MLE boundary solutions	MAP non-convergence	MAP boundary solutions	Start disagreement
Right-skewed	7,000	0	0	0	0	0	0
Symmetric	7,000	0	0	0	0	0	0
U-shaped	7,000	0	0	0	0	0	0
Total	21,000	0	0	0	0	0	0

5. Discussion & Practical Implications

The simulation results offer several insights that bridge classical statistics and modern data science. In our previous study, [1] established that the Method of Moments is preferable to MLE for small sample sizes because it is simpler and more numerically stable. Tables 1-3 are consistent with the accuracy side of that recommendation: at $N = 10$ MOM has a lower MSE than MLE in all three scenarios. The present results show, however, that the choice is not confined to the frequentist pair. For small samples, the Bayesian MAP estimator with the Gamma(2, 0.5) prior has a lower MSE than both MOM and MLE in the skewed and symmetric configurations, by a wide margin for $N \leq 20$ and by roughly 10-55% for $30 \leq N \leq 50$ (Tables 1 and 2). In the U-shaped configuration, the same prior does worse than MLE and MOM up to $N = 50$ (Table 3), so the advantage is not uniform across shapes.

The mechanism is straightforward. In small samples, the sample variance can fluctuate widely, which makes the MOM estimates occasionally very large (or, rarely, invalid), and the likelihood surface of the MLE is flat along the direction of large shape values, so that estimates are inflated and highly variable (Tables 1 and 2). The diagnostics of Section 4.5 show that this is not a failure of the optimizer: the optimizations converged in all cases. The Gamma prior adds a smooth penalty that pulls the estimates away from extreme values; this shrinkage reduces variance at the price of some bias. It pays off when the prior is centered reasonably near the true parameters, as in the skewed scenario at $N = 10$, where the MSE is reduced by roughly 90% (Table 1), and it costs accuracy when the prior is centered far from them, as in the U-shaped scenario (Table 3) and for the poorly located priors in Table 5.

The Bayesian MAP estimator also converges to the classical MLE. As the sample size increases ($N \geq 100$the contribution of the prior terms in the log-posterior diminishes, and by $N = 500$ the mean, bias and MSE of MAP and MLE are virtually identical in Tables 1, 2 and 3, and the MSE agrees to within about 7% for five of the six priors in Table 5 (the strongly informative Gamma(5, 1) prior remains up to about 17% above MLE in the symmetric and U-shaped scenarios). In the settings studied, a practitioner who uses MAP therefore pays no penalty in large samples. In small samples, however, the outcome depends on the prior, which is why we recommend that the sensitivity of the estimates to the hyperparameters is always examined, as done in Section 4.4.

This study has limitations. Only independent Gamma priors on the two shape parameters were considered, with six hyperparameter settings; MAP (the posterior mode, which is not invariant to re-parameterization) was studied rather than the posterior mean or fully Bayesian inference; three parameter configurations and 1,000 replications per condition were used, so differences of a few per cent between estimators should not be over-interpreted; and the convergence diagnostics refer to the specific optimizer configuration described in Section 3.1.

These trade-offs have direct, practical implications in industrial data science. Modern e-commerce and advertisement platforms frequently use Beta-Binomial models to estimate click-through rates (CTR) and user engagement. In these systems, estimating the Beta hyperparameters (α, β) from noisy and sparse click data is a core task (often called Empirical Bayes or rate smoothing, as seen at companies like Etsy).

Our research indicates that using a pure MLE or MOM to estimate these hyperparameters can lead to highly volatile results for new listings or items with few impressions. Assigning weak priors and utilizing a Bayesian MAP approach can give more stable CTR estimates, provided the prior is chosen with care, and so supports fairer and more robust content ranking and recommendation algorithms.

6. Conclusion

This study has presented a comparative analysis of the Method of Moments, Maximum Likelihood Estimation, and Bayesian Maximum A Posteriori estimators for the parameters of the Beta distribution. Our simulation results confirm and extend the classical findings of [1]. MOM remains a computationally trivial alternative for quick approximations, and MLE remains the standard for large-sample frequentist analysis. With the Gamma(2, 0.5) prior, Bayesian MAP achieves a substantial reduction in MSE in small samples when the prior is reasonably located (up to roughly 90% in the highly skewed case) and converges to MLE as the sample size increases. The sensitivity analysis shows, however, that this advantage is conditional: priors that are centered far from the true parameters give little gain or are worse than MLE and MOM, as the U-shaped scenario illustrates. Numerical diagnostics showed no optimizer failures or boundary solutions for MLE or MAP and no MOM variance violations in 21,000 datasets.

We therefore recommend that practitioners consider Bayesian MAP estimation in sparse-data applications such as CTR estimation, Thompson Sampling, and small-scale clinical trials when genuine prior information is available, that they choose the prior so that its mass covers the plausible range of the shape parameters (in particular not centered above small shape values), and that they report how the estimates change under alternative priors. When no such information exists, MOM or MLE, or a vague prior, are safer choices in small samples. Future research can explore fully Bayesian MCMC inference to capture the whole posterior distribution, empirical-Bayes choice of the hyperparameters, and further priors, such as Jeffreys' prior or non-informative flat priors.

References

[1] Akanihu, C. N., Hillary, A., & James, F. (2023). Investigating the Parameters of the Beta Distribution. World Journal of Advanced Research and Reviews, 19(1). https://doi.org/10.30574/wjarr.2023.19.1.1409.

[2] Gupta, A. K., & Nadarajah, S. (2004). Handbook of Beta Distribution and Its Applications. CRC Press. https://doi.org/10.1201/9781482276596.

[3] Aryafar, K., Guillory, C., & Halavanau, D. (2017). An Ensemble-based Approach to Click-Through Rate Prediction for Promoted Listings at Etsy. Proceedings of the AdKDD Workshop at KDD. https://doi.org/10.1145/3124749.3124758.

[4] Papadatos, N. D. (2024). On point estimators for Gamma and Beta distributions. The American Statistician, 78(4), 1–6. https://doi.org/10.1080/00031305.2024.2332766.

[5] Chen, P. & Xiao, X. (2025). Novel closed-form point estimators for the beta distribution. Statistical Theory and Related Fields, 9(1), 12-33. https://doi.org/10.1080/24754269.2024.2419360.

[6] Nawa, V. M., & Nadarajah, S. (2023). New Closed Form Estimators for the Beta Distribution. Mathematics, 11, 2799. https://doi.org/10.3390/math11132799.

[7] Tamae, H., Irie, K., & Kubokawa, T. (2020). A score-adjusted approach to closed-form estimators for the gamma and beta distributions. Japanese Journal of Statistics and Data Science, 3(2), 543–561. https://doi.org/10.1007/s42081-019-00071-x.