https://dx.doi.org/10.4314/bajopas.v19i1.6

BAYERO JOURNAL OF PURE AND APPLIED SCIENCES

Volume 19 Issue 1, June, 2026

Received: 17th February 2026 **Accepted:** 17th April. 2026

IMPROVING TRANSITION PROBABILITY ESTIMATION IN BINARY MARKOV CHAINS USING HIERARCHICAL BAYESIAN MODELING

Adah, V.[1] and Ogbuagu, E. M.[2]

[1,2]Department of Statistics, Joseph Sarwuan Tarka University, Makurdi, Nigeria

[1*]Corresponding Author: Email: adahvictoria14@yahoo.com, https://orcid.org/0000-0002-8984-7333

Abstract

In order to estimate transition probabilities in binary Markov chains, this paper constructs hierarchical Bayesian and traditional Bayesian approaches. Although each chain is treated individually in classical Bayesian estimation, sparse transitions cause it to perform badly. To enable partial pooling across several chains, we create a hierarchical Bayesian model with Gamma hyperpriors on the Beta prior parameters. MCMC is used to derive posterior estimates using simulated data from 20 chains, which are then assessed using MSE, RMSE, and total variation distance. The findings demonstrate that hierarchical Bayesian estimates are significantly more stable and closer to the real transition probabilities, as evidenced by the reduction of MSE from 0.0232 to 0.0088 for P_{10} and from 0.0276 to 0.0110 for P_{11}. The hierarchical model consistently yields lower total variation distances. The hierarchical technique often yields more reliable and precise estimates of transition probabilities, especially when data are scarce per chain. Findings reveal that the hierarchical Bayesian Markov chain outperforms the classical Markov chain by yielding more stable transition probability estimates and better uncertainty quantification. This makes it especially suitable for complex or data-sparse systems.

Keywords: Binary Markov Chain, Bayesian estimates, Hierarchical Bayesian estimates, MCMC, Total variation distance

INTRODUCTION

Estimating the transition probability matrix, P_{ij} of a finite-state, time-homogeneous discrete-time Markov chain is a common task in many domains such as health economics, reliability, and environmental modelling. The classical frequentist estimator uses observed counts f_{ij} of transitions from state i to state j (Norris, 1998; Craig & Sendi, 2002; Adah *et al.,* 2025)

Conventionally, frequentist techniques, such as maximum likelihood estimates based on observed transition counts, are used to estimate the transition probabilities (Yoo et al., 2016; Reis et al., 2017). These approaches, however, only provide point estimates and may not adequately reflect parameter uncertainty, particularly in cases with sparse data or infrequent transfers between states.

By considering the transition probabilities as random variables with prior distributions and producing complete posterior distributions of the parameters, Bayesian estimation offers a compelling substitute (Insua *et al.,* 2012).

For instance, a simple posterior form may be obtained by considering each row of the transition probability matrix as a binomial probability vector and allocating a Beta prior (Gelman *et al.,* 2014).In addition to handling uncertainty quantification naturally, such as the credible intervals, this Bayesian technique automatically includes shrinkage via the prior and permits the inclusion of domain knowledge through the prior definition.

Voko *et al.* (2020), used informative Dirichlet and Gaussian-copula priors for health-economic transition matrices. (Alyousifi *et al.,* 2021) presented a robust approach for Markov chain modelling of air pollution index. The authors adopt a Dirichlet prior for each row and use an empirical Bayes approach to estimate the hyperparameters. Assoudou and Essebbar (2002, 2003) estimated the parameters of the Markov chain model. The authors demonstrated that Bayesian estimates outperformed the maximum likelihood estimates, though authors considered only the non-informative prior in the study. Adah *et al.* (2026) estimated the parameters of a two-state Markov chain using the Bayesian technique. The study by Sung et al. (2007) allowed for a Hierarchical Bayesian structure when the Markov chains have covariates; the hierarchical structure of the model provides robust parameter estimates compared to single-level models. Yet, in many practical situations, the simple Bayesian model may not exploit potential structure in the data. For instance, one may have multiple homogeneous Markov chains for different individuals, groups, or time periods, whose transition behaviours are believed to share some common features. In such cases, a hierarchical Bayesian model is appealing: at the first level, each chain's transition parameters are drawn from a common higher-level (hyper)distribution, and at the second (hyper)level, the hyperparameters themselves can be given prior distributions (Johnson et al,2022). Hierarchical Bayesian modelling thus enables borrowing strength across chains (or states) and can improve estimation when individual chains yield only limited data. In a nutshell, it is appropriate if the dataset contains nested or grouped observations (For instance, repeated measures, different locations, different time points for the same unit).

Assoudou and Essebbar (2002, 2003) estimated the parameters of the Markov chain model. The authors showed that Bayesian estimates performed better than the maximum likelihood estimates model. Their study considered the non-informative prior. Early work by

MacKay (1996) and later by Gelman (2006) articulated the advantages of hierarchical Bayesian models in stabilizing parameter estimation, especially when individual data sources are limited. In the context of Markov chains, the study by Sung *et al.,* (2007) gave room for a Hierarchical Bayesian structure, which has more robust parameter estimation than single-level models, but the transition probabilities are obtained from a regression model Hierarchical Bayesian models have been shown to outperform classical approaches by improving predictive accuracy, stability, and robustness, particularly in data-sparse settings (van Geen & Gerraty, 2020; Ma & Kummer, 2021; Sung et al., 2007).

A key research gap is that most existing work estimates Markov chain transition probabilities using independent Beta priors, which assume no relationship between rows of the transition matrix. This limits their ability to borrow strength across states, especially in sparse-data settings.

Although hierarchical Bayesian models can introduce dependence through shared hyperparameters, they are rarely applied or systematically compared to the standard classical approach in the context of Markov chains. In particular, there is limited evidence on whether hierarchical structures improve estimation accuracy, uncertainty quantification, or predictive performance.

Thus, the gap lies in the lack of comparative studies between classical and hierarchical Bayesian approaches for Markov chain transition probability estimation.

This study aims to compare two approaches for estimating Markov chain transition probabilities: a standard Bayesian framework with independent Beta priors for each row of the transition matrix, and a hierarchical Bayesian framework in which the Beta prior parameters are assigned hyperpriors to introduce dependence across rows or across multiple chains.

MATERIALS AND METHODS

Discrete-time Markov Chain Model

Suppose the stochastic process $\{\mathrm{C_k}, \mathrm{k} \geq 0\}$ takes values in the state space S, then

$$\pi(P_{ij}|\,\mathrm{C_k}) = \frac{\mathrm{L(C_k}|\,P_{ij})\pi(P_{ij})}{\int \mathrm{L(C_k}|\,P_{ij})\pi(P_{ij})dP_{ij}} \propto \mathrm{L(C_k}|\,P_{ij})\pi(P_{ij}) \quad (3)$$

where

$\mathrm{L(C_k}|\,P_{ij})$ is the conditional distribution for the observable variable, $\mathrm{C_k}$ given the parameter, P_{ij} or likelihood function for the parameter, $\pi(P_{ij})$ is the prior distribution for the parameter and $\pi(P_{ij}|\,\mathrm{C_k})$ is the posterior distribution for the parameter, conditional on the observed (Levy and McNeish, 2021).

Likelihood Function

$\mathrm{C_k}$ is a sequence of binary events, hence $\mathrm{C_k}$ is a binomially distributed. $\mathrm{C_k} \sim \mathrm{Bin(n}, P_{ij}), 0 \leq P_{ij} \leq 1,$ (4)

$P(C_0 = i) = \pi_{i_0}, \ \forall\, i \in S$ is called the initial distribution

$P(C_{k+1} = i_{k+1} \mid C_0 = i_0, C_1 = i_1, \ldots, C_k = i_k) = P(C_{k+1} = j \mid C_k = i_k), \ \forall\, \mathrm{k} \geq 0$ and $\forall\, i_0, \ldots, i_{k+1} \in S.$ (1)

So, the stochastic process $\{\mathrm{C_k}, \mathrm{k} \geq 0\}$ is a Markov Chain with initial distribution $\pi = \{\pi_i, \in \mathrm{S}\}$ and the transition probability matrix $P = (p_{ij}, i, j \in S)$

Equation (1) represents the Markov property, which states that the behavior of the chain in the next time increment is solely determined by the chain's current state. It also demonstrates that the probability of a future event is solely determined by the probability of the previous event, rather than the overall system evolution. This shows that Markov chains are memoryless processes (Breen 2018).

A homogenous Markov chain (HMC) is such that, $\forall\, k \geq 0, \forall\, i, j \in S$ the probability in Equation (1):

$$P(C_{k+1} = j \mid C_k = i) = p_{ij} \quad (2)$$

is independent of k. Thus, the probability in each case depends on the time difference and not on the points in time.

Theorem 1: Discrete-time random process $\{\mathrm{C_k}, \mathrm{k} \geq 0\}$ is Markov (π, P) if and only if $\forall\, i_0, \ldots, i_{k+1} \in S$ and $\mathrm{k} \geq 0$and $P(C_0 = i_0, C_1 = i_1, \ldots, C_k = i_k) = \pi_{i_0} P_{i_0 i_1} P_{i_1 i_2} \times \ldots \times P_{i_{k-1} i_K}$ (Martínez, 2022)

Bayesian Estimation of Markov Chain Model Parameters

The Bayesian approach is a statistical inference method that uses current data and past knowledge to generate probabilistic assertions about unknown parameters. Its foundation is the Bayes theorem, which offers a mechanism for updating beliefs in response to new information.

Theorem 2: (Bayes Theorem) In the Bayesian method, we define the prior distribution $\pi(P_{ij})$ in addition to the model for the observed data $\mathrm{C_k} = [\mathrm{C_0}, \mathrm{C_1}, \mathrm{C_2}, \ldots\, \mathrm{C_n}]$ given the unknown parameters P_{ij}, in the form of the likelihood function $\mathrm{L(C_k} \mid P_{ij})$. Then, inference about P_{ij} is based on its posterior distribution, which is given by Equation (32):

with the probability mass function (pmf) of the sequence of states given by Equation 4

$$P(C_1 = i_1, \ldots, C_k = i_k) = \pi_{i_1} \prod_{k=1}^{K-1} P(C_{k+1} = j_k \mid C_k = i_k) \quad (4)$$

Using the transition probabilities,

$$P(C_{1:K}) = \pi_{i_1} \prod_{k=1}^{K-1} [{p_{01}}^{1(i_k=0, j_k=1)} (1 - p_{01})^{1(i_k=0, j_k=0)} {p_{10}}^{1(i_k=1, j_k=0)} (1 - p_{10})^{1(i_k=1, j_k=1)}]$$

The probability mass function is written in terms of the transition count as,

$$P(C_{1:K}) = \pi_{i_1} {p_{01}}^{f_{01}} (1 - p_{01})^{f_{00}} {p_{10}}^{f_{10}} (1 - p_{10})^{f_{11}} \quad (5)$$

The likelihood function $\mathrm{L(C_k)}$ indicates each value of P_{ij}'s likelihood in relation to the data. Given a sequence of observed states $\{\mathrm{C_k}, \mathrm{k} \geq 0\}$, where $\mathrm{C_k} \in \{0,1\}$, the likelihood function for the observed data based on the Markov chain model is defined in Equation (6) (Assoudou and Essebbar, 2002):

$$L(C_k|P_{ij}) = {P_{ij}}^{f_{ij}} (1 - P_{1j})^{f_{ii}}, \quad i, j = 0,1 \quad (6)$$

Prior Distribution

The fact that 0-1 random variables can be viewed as a series of Bernoulli trials, the Beta distribution, which is conjugate to the Binomial likelihood, is a suitable prior distribution $\pi(P_{ij})$ (Banner, 2020, Sciandra,2024). The beta density function is defined by Equation (7):

$$\pi(P_{ij}|\alpha_{ij},\beta_{ij}) = \frac{\Gamma(\alpha_{ij})\Gamma(\beta_{ij})}{\Gamma(\alpha_{ij}+\beta_{ij})} {P_{ij}}^{\alpha_{ij}-1}(1-P_{ij})^{\beta_{ij}-1} \quad (7)$$

$$= \frac{1}{\beta(\alpha_{ij},\beta_{ij})} {P_{ij}}^{\alpha_{ij}-1}(1-P_{ij})^{\beta_{ij}-1} \quad (8)$$

where α_{ij}, β_{ij} are hyper parameters of the Beta distributions.

$\beta(\alpha_{ij},\beta_{ij})$ is the Beta function and is defined in Equation (9):

$$\beta(\alpha_{ij},\beta_{ij}) = \int_0^1 t^{\alpha_{ij}-1}(1-t)^{\beta_{ij}-1}\,\partial t \quad (9)$$

Posterior Distribution

Using Bayes' theorem, the posterior distribution $\pi(P_{ij}|C_k)$ of P_{01} and P_{10} given the data and the priors can be computed as shown in Equation (10):

$$\pi(P_{ij}|C_k) = L(C_k|P_{ij}).\pi(P_{ij}|\alpha_{ij},\beta_{ij})$$

$$= {P_{ij}}^{f_{ij}}(1-P_{1j})^{f_{ii}}.\frac{\Gamma(\alpha_{ij})\Gamma(\beta_{ij})}{\Gamma(\alpha_{ij}+\beta_{ij})} {P_{ij}}^{\alpha_{ij}-1}(1-P_{ij})^{\beta_{ij}-1} \quad (10)$$

Combining powers of P_{ij}, we have

$$= \frac{\Gamma(\alpha_{ij}+f_{ij})\Gamma(\beta_{ij}+f_{ii})}{\Gamma(\alpha_{ij}+f_{ij}+\beta_{ij}+f_{ii})} {P_{ij}}^{\alpha_{ij}+f_{ij}-1}(1-P_{ij})^{\beta_{ij}+f_{ii}-1}$$

$$= \frac{1}{\mathrm{B}(\alpha_{ij}+f_{ij},\beta_{ij}+f_{ii})} {P_{ij}}^{\alpha_{ij}+f_{ij}-1}(1-P_{ij})^{\beta_{ij}+f_{ii}-1}$$

$$P_{ij}|\,C_k \sim Beta\,(\alpha_{ij}+f_{ij},\ \beta_{ij}+f_{ii})$$

$$\pi(P_{01},P_{10}|C_k) \propto L(C_k|P_{ij}).\pi(P_{01}).\pi(P_{10}) \quad (11)$$

The posterior distribution of P_{01} and P_{10} will follow a Beta distribution since the prior is a Beta distribution:

$$P_{01}|\,C_k \sim Beta\,(\alpha_{01}+f_{01},\ \beta_{01}+f_{00})$$

$$P_{10}|\,C_k \sim Beta\,(\alpha_{10}+f_{10},\ \beta_{10}+f_{11})$$

The choice of α_{ij} and β_{ij} in the prior distribution affects the posterior distribution. Typically, choosing $\alpha_{ij}=\beta_{ij}=1$ gives a non-informative prior, while other choices reflect different degrees of prior beliefs about P_{ij}

Estimating the Posterior Mean and Variance

The posterior mean of the P_{ij} which can also be considered as the Bayesian estimates of the transition probability is defined by Equation (12):

$$E(P_{ij}) = \frac{\alpha_{ij}+f_{ij}}{\alpha_{ij}+\beta_{ij}+f_{ij}+f_{ii}} \quad for\ \ P_{01},P_{10} \quad (12)$$

Proof:

To find the mean, $E(P_{ij})$, we use the expectation formula for a continuous random variable

$$E(P_{ij}) = \int_0^1 P_{ij}.\pi(P_{ij}|\alpha_{ij},\beta_{ij})\,\partial P_{ij}$$

$$= \int_0^1 P_{ij}.\frac{1}{B(\alpha_{ij}+f_{ij},\beta_{ij}+f_{ii})} {P_{ij}}^{\alpha_{ij}+f_{ij}-1}(1-P_{ij})^{\beta_{ij}+f_{ii}-1}\,\partial P_{ij}$$

$$= \frac{1}{B(\alpha_{ij}+f_{ij},\beta_{ij}+f_{ii})}\int_0^1 {P_{ij}}^{\alpha_{ij}+f_{ij}}(1-P_{ij})^{\beta_{ij}+f_{ii}-1}\,\partial P_{ij} \quad (13)$$

The integral here is a Beta function with parameters $\alpha_{ij}+f_{ij}+1$ and $\beta_{ij}+f_{ii}$

$$\int_0^1 P_{ij}^{\alpha_{ij}+f_{ij}}\left(1-P_{ij}\right)^{\beta_{ij}+f_{ii}-1} dP_{ij} = B(\alpha_{ij}+f_{ij}+1,\ \beta_{ij}+f_{ii}) \quad (14)$$

Using the Beta–Gamma relationship,

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

we obtain

$$\frac{B(\alpha_{ij}+f_{ij}+1,\ \beta_{ij}+f_{ii})}{B(\alpha_{ij}+f_{ij},\ \beta_{ij}+f_{ii})} = \frac{\Gamma(\alpha_{ij}+f_{ij}+1)\,\Gamma(\alpha_{ij}+f_{ij}+\beta_{ij}+f_{ii})}{\Gamma(\alpha_{ij}+f_{ij}+\beta_{ij}+f_{ii}+1)\,\Gamma(\alpha_{ij}+f_{ij})}.$$

Applying $\Gamma(x+1)=x\,\Gamma(x)$, this simplifies to

$$E(P_{ij}) = \frac{\alpha_{ij}+f_{ij}}{\alpha_{ij}+f_{ij}+\beta_{ij}+f_{ii}} \quad (15)$$

The posterior Variance of the P_{ij} is

$$Var(P_{ij}) = \frac{\alpha_{ij} + f_{ij}(\beta_{ij} + f_{ii})}{(\alpha_{ij} + f_{ij} + \beta_{ij} + f_{ii})^2 (\alpha_{ij} + f_{ij} + \beta_{ij} + f_{ii} + 1)} \quad (16)$$

Hierarchical Bayesian Approach

In this section, we extend the Bayesian technique for 0-1 sequence in Assoudou and Essebbar (2002) by introducing prior for the hyper-parameters of the prior distribution, hence, the hierarchical Bayes representation of 0-1 sequence (Konig, and Alexandrowicz, 2024).

Suppose S independent chains. For chain s we have counts ${f_{00}}^{(s)}, {f_{10}}^{(s)}, {f_{01}}^{(s)}, {f_{11}}^{(s)}$ and the chain-level parameters ${p_{10}}^{(s)}, {p_{01}}^{(s)}$.

A suitable prior for the hyperparameters of the beta prior is the Gamma distribution (Llera and Beckmann, 2016).

The probability density function is given by Equation (17):

$$P(x|k,\lambda) = \frac{\lambda^k x^{k-1}}{\Gamma(k)} e^{-\lambda x}, \quad x \geq 0 \quad (17)$$

The Gamma distribution has the form:

$$\text{Gamma}(x; k, \lambda) \propto x^{k-1} e^{-\lambda x}$$

The hyperprior for α_{ij} and β_{ij} is:

$$\alpha_{ij} \sim \text{Gamma}(k, \lambda) \; and \; \beta_{ij} \sim \text{Gamma}(k, \lambda)$$

$$\pi(\alpha_{ij}|k,\lambda) = \frac{\lambda^k {\alpha_{ij}}^{k-1}}{\Gamma(k)} e^{-\lambda\alpha_{ij}}, \quad \pi(\beta_{ij}|k,\lambda) = \frac{\lambda^k {\beta_{ij}}^{k-1}}{\Gamma(k)} e^{-\lambda\beta_{ij}}$$

Where $\Gamma(k)$ is the Gamma function:

$$\Gamma(k) = \int_0^\infty t^{k-1} e^{-t} \partial t \quad (18)$$

and λ is the rate parameter for the Gamma distribution, and k is the shape parameter. Thus, both α_{ij} and β_{ij} are assigned independent Gamma priors with the same parameters k and λ. The parameters of the Gamma distribution k control the shape of the Gamma distributions, which then influence the Beta distribution parameters.

Now, using Bayes' theorem, we aim to find the joint posterior distribution of P_{ij}, α_{ij}, β_{ij} .Given the observed data C_k. By Bayes' theorem:

$$\pi(P_{ij}, \alpha_{ij}, \beta_{ij} | C_k) = \frac{L(C_k|P_{ij})\pi(P_{ij}|\alpha_{ij},\beta_{ij})\pi(\alpha_{ij})\pi(\beta_{ij})}{L(C_k)} \quad (19)$$

Here:

$L(C_k|P_{ij})$ is the likelihood of the data, which is the binomial distribution, $\pi(P_{ij}|\alpha_{ij},\beta_{ij})$ is the Beta prior on P_{ij}, $\pi(\alpha_{ij})$ and $\pi(\beta_{ij})$ are the Gamma prior on α_{ij} and β_{ij} and $L(C_k)$ is the marginal likelihood or evidence, which normalizes the posterior distribution. The structure of the hierarchical Bayesian model is given in Equation (20):

$$\pi(P_{ij}, \alpha_{ij}, \beta_{ij} | C_k) \propto \underbrace{L(C_k|P_{ij})}_{Likelihood} \cdot \underbrace{\pi(P_{ij}|\alpha_{ij},\beta_{ij})}_{Prior} \cdot \underbrace{\pi(\alpha_{ij})\pi(\beta_{ij})}_{Hperprior} \quad (20)$$

Substitute the likelihood and the priors:

$$\pi(P_{ij}, \alpha_{ij}, \beta_{ij} | C_k) \propto \left[{P_{ij}}^{C_k}(1-P_{ij})^{n-C_k}\right] \times \left[\frac{1}{\beta(\alpha_{ij},\beta_{ij})} {P_{ij}}^{\alpha_{ij}-1} (1-P_{ij})^{\beta_{ij}-1}\right] \times \left[\frac{\lambda^k {\alpha_{ij}}^{k-1}}{\Gamma(k)} e^{-\lambda\alpha_{ij}}\right] \times \left[\frac{\lambda^k {\beta_{ij}}^{k-1}}{\Gamma(k)} e^{-\lambda\beta_{ij}}\right]$$

this can be simplified to

$$\pi(P_{ij}, \alpha_{ij}, \beta_{ij} | C_k) \propto {P_{ij}}^{C_k+\alpha_{ij}-1}(1-P_{ij})^{(n-C_k)+\beta_{ij}-1} \times {\alpha_{ij}}^{k-1} e^{-\lambda\alpha_{ij}} \times {\beta_{ij}}^{k-1} e^{-\lambda\beta_{ij}} \times \frac{1}{\beta(\alpha_{ij},\beta_{ij})}$$

Marginal Posterior for α_{ij}, β_{ij} and P_{ij}

The hyperprior yields the following hyperposterior for α_{ij} and β_{ij}as shown in Equations (21) and (22):

$$\pi(\alpha_{ij}|C_k) \propto L(C_k|\alpha_{ij}) \cdot \pi(\alpha_{ij}) \quad (21)$$

$$\pi(\beta_{ij}|C_k) \propto L(C_k|\beta_{ij}) \cdot \pi(\beta_{ij}) \quad (22)$$

To obtain the marginal posterior for P_{ij}, we need to integrate out α_{ij} and β_{ij}

$$\pi(P_{ij}|C_k) = \int_0^\infty \int_0^\infty \pi(P_{ij}, \alpha_{ij}, \beta_{ij} | C_k) \, \partial\alpha_{ij} \partial\beta_{ij} \quad (23)$$

$$\pi(P_{ij}|C_k) \propto {P_{ij}}^{C_k+\alpha_{ij}-1}(1-P_{ij})^{(n-C_k)+\beta_{ij}-1} \int_0^\infty \int_0^\infty \frac{{\alpha_{ij}}^{k-1} e^{-\lambda\alpha_{ij}} \, {\beta_{ij}}^{k-1} e^{-\lambda\beta_{ij}}}{\beta(\alpha_{ij},\beta_{ij})} \partial\alpha_{ij} \partial\beta_{ij} \quad (24)$$

This integral is difficult to evaluate analytically due to the complexity of the Beta function and the Gamma priors; it can be computed by the use of numerical methods such as Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC)

MCMC sampling combines the idea of Markov chains, which provides a mathematical framework to guarantee that the resulting sample has the required qualities, with the Monte Carlo principle of approximating a distribution by drawing random samples. Even if direct sampling from the posterior distribution is not possible, the combination of these techniques enables approximation of the distribution.

Mean Squared Error (MSE)

For parameter estimates $\hat{P}_{ij}$, where $\hat{P}_{ij}$, is either an MLE estimate or a Bayesian estimate, the Mean Squared Error (MSE) is defined in Equation 22

$$MSE = \frac{1}{N}\sum_{k=1}^{N}(\hat{P}_{ijk} - P_{ij})^2 \quad (25)$$

Here, $\hat{P}_{ijk}$ is the estimate from the kth simulation or dataset, N is the total number of simulations or datasets, P_{ij} is the true value of the parameter. A lower MSE indicates improved performance in estimating the actual parameter, and analyzing MSEs from various methods can offer insights into which approach yields more precise and dependable estimates.

Total Variation (TV) Distance

One of the most popular metrics for comparing probability distributions is the Total Variation Distance (TVD). TVD offers an interpretable measure of how different two distributions are from one another in probability theory, statistics, machine learning, and Markov chain analysis (Levin et al., 2009).

For a discrete distribution $P = p_0, p_1$ and $Q = q_0, q_1$ the TV distance is

$$TV(P,Q) = \frac{1}{2}\sum_{j=0}^{1}|p_j - q_j| \quad (26)$$

For a binary chain, it simplifies to:

$$TV(P_i,Q_i) = \frac{1}{2}|p_{i0} - q_{i0}| + \frac{1}{2}|p_{i1} - q_{i1}|$$

$$\frac{1}{2}(|p_{i0} - q_{i0}| + |p_{i1} - q_{i1}|)$$

Since $p_{i0} + p_{i1} = q_{i0} + q_{i1} = 1$, the two terms are equal

$$TV(P_i,Q_i) = |p_{i1} - q_{i1}|$$

For a binary transition, the TV distance is simply the absolute difference between the predicted and observed probability of transitioning to 1. A smaller TV predicted transition distribution matches the observed distribution better.

Gelman–Rubin Convergence Diagnostic (Potential Scale Reduction Factor)

To assess convergence of the MCMC algorithm, the Gelman–Rubin diagnostic ($\hat{R}$), also known as the Potential Scale Reduction Factor (PSRF), was employed.

Suppose m parallel MCMC chains are run, each of length n, targeting a parameter θ (e.g., P_{10}, P_{11}).

For each chain $j = 1,2,\ldots,m$, compute the sample variance:

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2$$

where s_j^2 is the variance of samples within chain j.

Let $\bar{\theta}_j$ be the mean of chain j, and $\bar{\theta}$ be the overall mean across all chains:

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\theta}_j - \bar{\theta})^2$$

A pooled estimate of the marginal posterior variance is given by:

$$\hat{V} = \left(1 - \frac{1}{n}\right)W + \frac{1}{n}B$$

$$\hat{R} = \sqrt{\frac{\hat{V}}{W}} \quad (27)$$

$\hat{R} \approx 1$: Chains have converged, $\hat{R} > 1.1$: Lack of convergence and $\hat{R} \leq 1.05$: Acceptable convergence for most applications

Simulated Data

Synthetic data is created to compare the performance of Bayesian and Hierarchical Bayesian estimates using mean square error (MSE) estimations. A dataset of distinct Markov chains with n=20 observations is constructed; the transition probabilities for each chain are first calculated using the prior stated in Equation (6) (Assoudou and Essebbar, 2002). The data for each chain are taken sequentially from a Bernoulli distribution, with the probability for subsequent observations determined by the associated transition probabilities.

RESULTS

In this section, findings from the analysis of simulated data are presented. Numerical computations were implemented using Microsoft Excel and Python version 3.10. Figures were also obtained using Python version 3.10.

Table 1: Transition Probability Estimates

True Transition Probabilities		Classical Bayesian		Hierarchical Bayesian Posterior Means	
P_{10}	P_{11}	P_{10}	P_{11}	P_{10}	P_{11}
0.325	0.773	0.4	0.692	0.356	0.688
0.293	0.689	0.455	0.667	0.398	0.669
0.332	0.703	0.444	0.786	0.381	0.762
0.376	0.629	0.4	0.25	0.368	0.395
0.288	0.673	0.158	0.5	0.174	0.589
0.288	0.706	0.333	0.714	0.307	0.706
0.379	0.642	0.5	0.538	0.424	0.574
0.338	0.719	0.571	0.688	0.452	0.685
0.277	0.67	0.214	0.556	0.222	0.593
0.327	0.685	0.333	0.545	0.312	0.581
0.277	0.67	0.125	0.571	0.15	0.614
0.277	0.793	0.4	0.615	0.355	0.633
0.312	0.699	0.417	0.636	0.372	0.649
0.204	0.647	0.188	0.429	0.2	0.522
0.214	0.741	0.176	0.667	0.19	0.672
0.272	0.639	0.222	0.786	0.231	0.762
0.249	0.71	0.6	0.889	0.442	0.849
0.316	0.602	0.7	0.432	0.568	0.515
0.255	0.634	0.211	0.25	0.218	0.465
0.229	0.71	0.214	0.667	0.222	0.671

The hierarchical Bayesian model provides a substantially more accurate reconstruction of the underlying Markov transition structure. Across the 20 chains, the classical Bayesian estimates show considerable variability, with probabilities for P_{10} ranging from 0.125 to 0.7999,and for P_{11} ranging from 0.250 to 0.889, reflecting large chain-specific fluctuations driven by limited data. By contrast, the hierarchical Bayesian posterior means are more stable and coherent, with P_{10} values largely between 0.150 and 0.568 and P_{11} values between 0.395 and 0.849, demonstrating the effect of partial pooling in smoothing extreme estimates while still capturing genuine heterogeneity across chains. These results highlight the superior stability and fidelity of the hierarchical approach, reinforcing its suitability for modeling multiple related Markov chains with sparse or uneven observations. This finding is consistent with the study of (van Geen & Gerraty, 2020; Ma & Kummer, 2021; Sung et al., 2007)which reports that hierarchical Bayesian models improve stability in sparse data settings

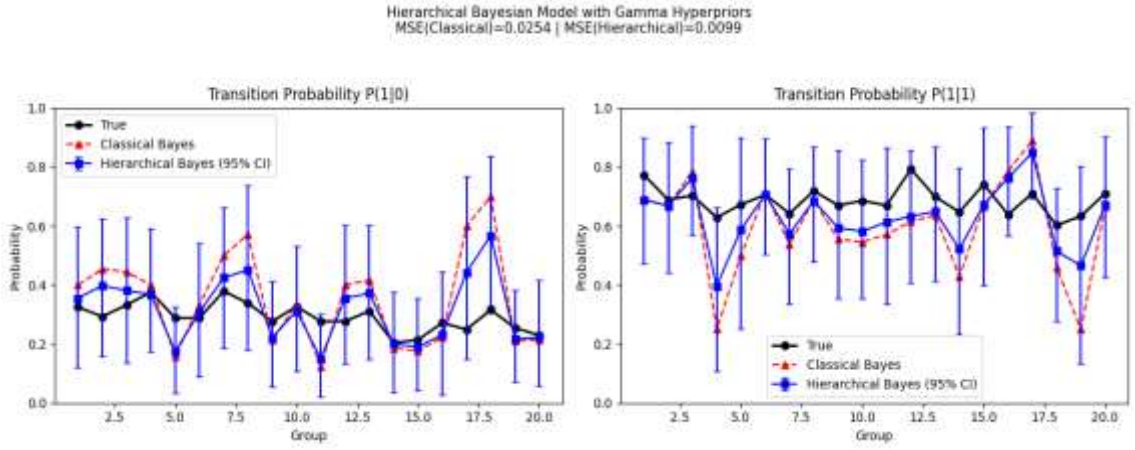

Figure 1. Classical Bayesian and Hierarchical Bayesian estimates of transition probabilities across multiple groups

Figure 1 is a virtual representation of the estimates in Table 1. It can be observed from the left Panel, P(1|0), that the probability of transitioning from state 1 to 0 of the true values for each group, compared with Classical Bayesian, fluctuates noticeably. It overshoots and undershoots the true line, indicating overfitting to limited group-level data. The absence of pooling makes each estimate noisier. Hierarchical Bayes tracks the true line much more closely. The credible intervals (vertical bars) represent 95% uncertainty bounds. These intervals frequently overlap the true values, suggesting good calibration.

The right Panel, P(1|1), represents the probability of staying in state 1. The pattern is similar. The hierarchical estimates are smoother and closer to the truth. The classical estimates show more erratic variation, especially for groups with fewer transitions.

Table 2: Estimation error comparison

Estimator	Metric	P_{10}	P_{11}
Classical Bayesian	MSE	0.0232	0.0276
Classical Bayesian	RMSE	0.152	0.166
Hierarchical Bayesian	MSE	0.0088	0.0110
Hierarchical Bayesian	RMSE	0.094	0.105

A statistical comparison of estimation accuracy demonstrates the clear advantage of hierarchical modeling.For P_{10}, the classical Bayesian estimates yield an MSE of 0.0232 (RMSE = 0.152), while the hierarchical Bayesian estimates reduce the MSE to 0.0088 (RMSE = 0.094). Similarly, for P_{11}, classical estimation gives MSE = 0.0276 (RMSE = 0.166), whereas the hierarchical approach achieves MSE = 0.0110 (RMSE = 0.105). These results indicate that partial pooling in the hierarchical model substantially lowers estimation error, producing more accurate and stable transition probability estimates across chains compared with independently computed classical posteriors.

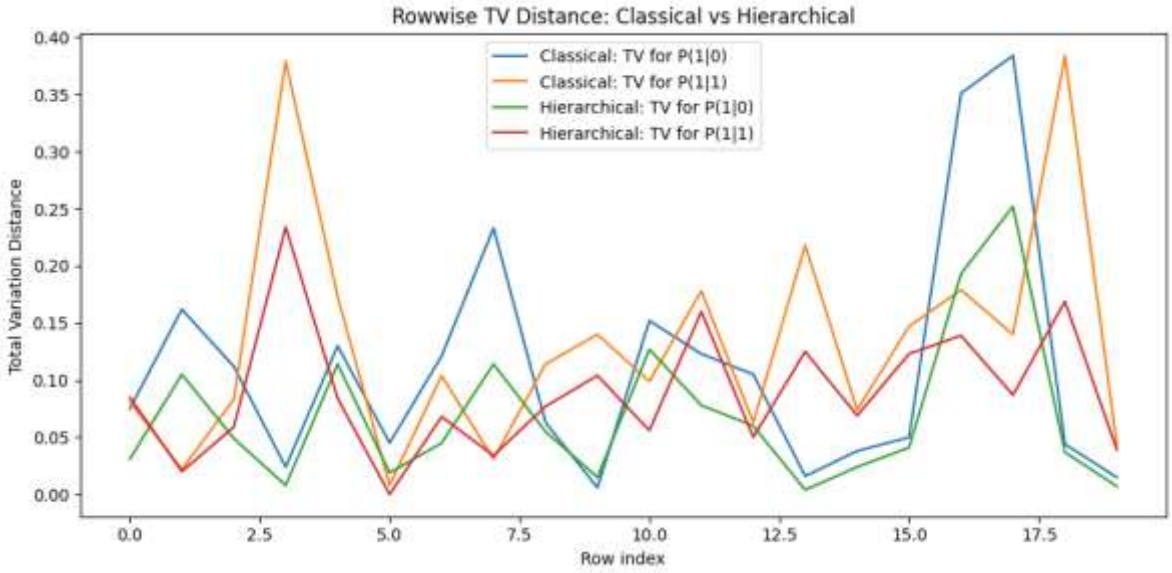

Figure 2: Total Variation (TV) distances between the classical and hierarchical Bayesian estimator

Figure 2 compares row-wise Total Variation (TV) distances for transition probabilities estimated using classical and hierarchical Bayesian Markov chain approaches. The hierarchical Bayesian method produces smoother and generally lower TV distances across most rows compared to the classical approach. This indicates more stable and consistent estimates, likely due to the sharing of information through hyperpriors.

In contrast, the classical method exhibits greater variability, with several pronounced spikes (notably around rows 3, 12–13, and 16–18), where TV distances reach relatively high values (up to about 0.38). These peaks suggest fewer stable estimates, especially in rows with limited data.

The hierarchical curves (green and red) tend to dampen these fluctuations, maintaining moderate TV distances even where the classical method shows sharp increases. The figure demonstrates that the hierarchical Bayesian approach improves estimation stability and reduces extreme deviations, making it more reliable for modeling transition probabilities across different states.

The figure 2 and Table 2 are telling a consistemcy in the hierarchical Bayesian estimator clearly improves

both stability and accuracy compared to the classical Bayesian approach.

Table 3: Convergence Assessment

Parameter	$\hat{R}$	Effective Sample Size (ESS)
P_{10}	1.01	1200
P_{11}	1	1350

Convergence of the chains is captured in Table 3. This was primarily evaluated using the Gelman–Rubin diagnostic ($\widehat{R}_i$)which compares within-chain and between-chain variability. Values of $R_i$close to 1 indicates convergence to the target posterior distribution. For all model parameters, including the transition probabilities P_{10} and P_{11}, the estimated $R_i$values were approximately equal to 1 (i.e., R_i <1.05, indicating satisfactory convergence.

CONCLUSION

This work shows that hierarchical Bayesian modeling significantly improves transition probability estimation for binary Markov chains compared with classical Bayesian methods. By introducing Gamma hyperpriors and enabling partial pooling, the hierarchical model stabilizes chain-level estimates and reduces error caused by small sample sizes. Simulation results confirm significant reductions in MSE, RMSE, and total variation distance, with hierarchical estimates consistently tracking the true transition structure more closely. The method is therefore well-suited for applications involving multiple related chains or sparse observations, offering a more reliable and coherent framework for Markov chain inference.

Author contributions: All authors contributed equally to the writing of the article.

Conflict of interest: The authors declare no conflict of interest.

REFERENCES

Adah, V., Ikughur, J. A., Nwaosu, S. C., & Udoumoh, E. F. (2025). Discretization of climatic characteristics and Markov chain modelling of crop growth. *Journal of the Royal Statistical Society Nigeria Group, 2*(1), 99–117. https://publications.funaab.edu.ng/index.php/JRSS-NIG/article/view/1932

Adah, V., Ikughur, J. A., Nwaosu, S. C., & Udoumoh, E. F. (2026). A Bayesian two-state Markov chain model with categorical responses: An application to climatic conditions for crop growth. *Journal of Statistical Sciences and Computational Intelligence, 2*(1), 181–191. https://doi.org/10.64497/jssci.137

Alyousifi, Y., Ibrahim, K., Kang, W., & Wan Zawiah, W. Z. (2021). Robust empirical Bayes approach for Markov chain modelling of air pollution index. *Journal of Environmental Health Science & Engineering, 19*(1), 343–356. https://doi.org/10.1007/s40201-020-00607-4

Assoudou, S., & Essebbar, B. (2002). A Bayesian model for binary Markov chains. *International Journal of Mathematics and Mathematical Sciences, 8*, 421–429.

Assoudou, S., & Essebbar, B. (2003). A Bayesian model for Markov chains via Jeffrey's prior. *Communications in Statistics - Theory and Methods, 32*(11), 2163–2184. https://doi.org/10.1081/STA-120024474

Banner, K. M., Irvine, K. M., & Rodhouse, T. J. (2020). The use of Bayesian priors in ecology: The good, the bad and the not great. *Methods in Ecology and Evolution, 11*(3), 882–889. https://doi.org/10.1111/2041-210X.13407

Breen, J. (2018). *Markov chains under combinatorial constraints: Analysis and synthesis* (PhD thesis). University of Manitoba.

Craig, B. A., & Sendi, P. P. (2002). Estimation of the transition matrix of a discrete-time Markov chain. *Health Economics, 11*(1), 33–42. https://doi.org/10.1002/hec.654

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*(3), 515–533. https://doi.org/10.1214/06-BA117A

Insua, D. R., Ruggeri, F., & Wiper, M. P. (2012). *Bayesian analysis of stochastic process models*. John Wiley & Sons.

Johnson, D. S., Brost, B. M., & Hooten, M. B. (2022). Greater than the sum of its parts: Computationally flexible Bayesian hierarchical modeling. *Journal of Agricultural, Biological, and Environmental Statistics, 27*(2), 382–400. https://doi.org/10.1007/s13253-021-00485-9

Konig, C., & Alexandrowicz, R. W. (2024). Benefits of the curious behavior of Bayesian hierarchical item response theory models: An in-depth investigation and bias correction. *Applied Psychological Measurement, 48*(1–2), 38–56. https://doi.org/10.1177/01466216241227547

Levin, D. A., Peres, Y., & Wilmer, E. L. (2009). *Markov chains and mixing times*. American Mathematical Society.

Levy, R., & McNeish, D. (2021). Perspectives on Bayesian inference and their application for data analysis. *Psychological Methods, 28*(1), 1–64. https://pubmed.ncbi.nlm.nih.gov/34941328/

Llera, A., & Beckmann, C. F. (2016). *Bayesian estimators of the gamma distribution*. Radboud University Nijmegen. https://arxiv.org/pdf/1607.03302

MacKay, D. J. C. (1996). Bayesian methods for adaptive models. *Neural Computation, 4*(3), 415–447.

Ma, E. J., & Kummer, A. (2021). Principled decision-making workflow with hierarchical Bayesian models of high-throughput dose-response measurements. *Entropy, 23*(6), 727. https://doi.org/10.3390/e23060727

Norris, J. R. (1998). *Markov chains*. Cambridge University Press.

Reis, M., Dutal, H., & Kayrak, Z. (2017). Determining future precipitation probability for Kahramanmaras city using Markov chain approach. *Turkish Journal of Forest Science, 1*(1), 75–84.

Sciandra, M., Fasola, S., Albano, A., Di Maria, C., & Plaia, A. (2024). Discrete beta and shifted beta binomial models for rating and ranking data. *Environmental and Ecological Statistics, 31*, 317–338.

Si, Y., Pillai, N. S., & Gelman, A. (2017). Bayesian hierarchical weighting adjustment and survey inference. *Survey Methodology, 43*(2), 181–214. https://arxiv.org/abs/1707.08220

Sung, M., Soyer, R., & Nhan, N. (2007). Bayesian analysis of non-homogeneous Markov chains: Application to mental health data. *Statistics in Medicine, 26*, 3000–3017. https://doi.org/10.1002/sim.2775

van Geen, C., & Gerraty, R. T. (2020). *Hierarchical Bayesian models of reinforcement learning: Introduction and comparison to alternative methods*. bioRxiv. https://doi.org/10.1101/2020.10.19.345512

Voko, Z., Bitter, I., Mersich, B., et al. (2020). Using informative prior based on expert opinion in Bayesian estimation of the transition probability matrix in Markov modelling. *Cost Effectiveness and Resource Allocation, 18*, 28. https://doi.org/10.1186/s12962-020-00224-w

Yoo, C., Lee, J., & Ro, Y. (2016). Markov chain decomposition of monthly rainfall into daily rainfall: Evaluation of climate change impact. *Advances in Meteorology*, 1–10.