

Inferring Transition Dynamics from Histogram Time Series

Inferring Transition Dynamics from Histogram Time Series: Identifiability at and Near Equilibrium

Ido Nevat

idonevat@gmail.com

Nanyang Technological University

Research Article

Keywords: aggregate Markov data, continuous-time Markov chains, histogram time series, structural identifiability, reversible Markov chains, Gaussian approximation

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11226228/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Inferring Transition Dynamics from Histogram Time Series: Identifiability at and Near Equilibrium

Ido Nevat

College of Computing and Data Science, Nanyang Technological University, Singapore
Earth Observatory of Singapore, Nanyang Technological University, Singapore

Abstract

Repeated histograms record the composition of a persistent population while concealing individual transitions. We study independent units evolving under a common finite-state continuous-time Markov chain and ask which features of its transition dynamics remain identifiable from these aggregate observations. Our central result concerns the destination covariance conditional on a source histogram fixed at the equilibrium composition. For known population size and observation interval, this covariance and the equilibrium composition uniquely determine the transition matrix and generator within the class of reversible continuous-time chains. The inversion uses a principal matrix square root followed by a logarithm. It complements recovery from stationary lagged cross-covariance by using fluctuations conditional on a fixed source rather than dependence between random source and destination histograms. We also give sufficient conditions for identifying the labeled transition matrix from aggregate conditional laws. An exact finite-population likelihood, evaluated by dynamic programming, and a Gaussian approximation provide computational tools for inference. Within the conditional Gaussian model, we characterize information about parameter changes that preserve the stationary distribution. Along source compositions approaching equilibrium as $N^{-\alpha}$, the mean contribution scales as $N^{1-2\alpha}$ in nondegenerate directions, while the covariance contribution has a finite limit. Their crossover occurs at $N^{-1/2}$ when both limiting sensitivities are nonzero. Numerical studies compare exact and Gaussian likelihood fits, illustrate conditional-covariance recovery, and examine population size, observation length, and temporal resolution. The results distinguish population-level identification from finite-sample recovery and show how equilibrium fluctuations can retain information about hidden transition dynamics.

Keywords: aggregate Markov data; continuous-time Markov chains; histogram time series; structural identifiability; reversible Markov chains; Gaussian approximation.

MSC 2020: 60J27; 62M05; 60J22.

1 Introduction

Many systems are observed repeatedly through counts rather than individual trajectories. At each observation time, we may know how many members of a population occupy each state, while the identities linking one observation to the next are unavailable. The resulting data form a time series of histograms. Such observations describe population composition, but they do not directly reveal the movements that produced it.

This distinction matters because composition and transition activity are not the same thing. A histogram can remain almost unchanged because little movement occurred, or because substantial movement in opposite directions happened to balance out. The inferential problem is therefore not simply whether the population distribution changes over time, but what repeated aggregate observations reveal about the transition dynamics operating underneath that distribution.

Core Research Question:

What can be learned about individual transition dynamics when only aggregate histograms are observed over time?

The question is especially revealing at equilibrium. When the population composition is stable, the expected histogram no longer changes from one observation to the next, even though individuals may continue to move actively between states. If the mean composition is no longer informative about transition activity, any remaining information must appear elsewhere in the aggregate observations. This is the central issue studied in this paper.

Inference from aggregate Markov data has a substantial history. Voting data provide a familiar example: successive marginal distributions of choices may be available while voter-level transitions remain unobserved, leading to ecological-inference problems (Brown and Payne, 1986; King, 1997). Related multistate problems arise in longitudinal and demographic studies, including estimation of transition probabilities from aggregate samples (Hawkins and Han, 2000), competing-risks and mortality applications (Davis et al., 2002), and health-expectancy analyses based on counts in disability states (Heathcote and Puza, 2005). Kalbfleisch et al. (1983) studied continuous-time Markov populations observed through state totals; Kalbfleisch and Lawless (1984) considered discrete-time transition estimation from successive aggregate counts; and Lawless and McLeish (1984) examined the information retained after individual transition histories are removed. Crowder and Stephens (2011) later developed transform-based inference for Markov-chain macro-data. These works establish repeated state-count data as an important form of incomplete Markov observation.

Stationary aggregate observations can also identify transition dynamics through moments. Bernstein and Sheldon (2016) develop moment-based estimation from aggregate observations, including noisy counts, using the relation between aggregate moments and individual-state marginals. Our equilibrium result considers a different second-order observable: the covariance of the destination histogram conditional on the source histogram being fixed at the equilibrium composition. Under reversible continuous-time dynamics, we show that this conditional covariance, together with the equilibrium composition, can be inverted constructively to recover the finite-time transition matrix and its generator. Section 4.3 relates this result explicitly to stationary lagged-covariance recovery.

Table 1 summarizes the observables and assumptions underlying these different conclusions. The reversible embedding result of Jia (2016) supplies relevant matrix-logarithm theory; the equilibrium result here starts one step earlier, with a conditional covariance that determines the transition matrix itself. The near-equilibrium analysis applies the standard Gaussian Fisher-information formula to quantify sensitivity as the source composition approaches equilibrium.

More recent work approaches incomplete or indirect observation of transition mechanisms from several directions. Mascherpa et al. (2026) formulate estimation of a common transition matrix and latent transport plans as an inverse optimal-transport problem for aggregate snapshots and time series. Baltazar-Larios and Esparza (2022) develop likelihood-based inference for a partially observed Markov-modulated process. Proportion-valued time series can instead be modeled directly through latent-state formulations; Can et al. (2022), for example, use a hidden Markov model for proportions. In the present model, the histograms arise by aggregating a persistent population whose individual members evolve independently according to a common continuous-time Markov transition mechanism. We focus on what remains learnable about this mechanism after individual correspondences have been lost. The likelihood and structural-identifiability analysis do not require reversibility or a particular transition graph. Additional structure is introduced only in the equilibrium analysis where it is needed.

The loss of unit correspondences also creates a computational problem. The likelihood of two successive histograms must account for all unobserved transition tables that are compatible with their margins. Direct enumeration quickly becomes impractical. We therefore develop an exact

Table 1: Information available from aggregate observations. The recovery statements assume that the indicated population moments or laws are known; they are not finite-sample accuracy guarantees.

Setting	Observable and assumptions	Conclusion
Stationary aggregate moments	Stationary composition and lag-one cross-covariance; independent units; positive stationary probabilities	Recover $\mathbf{P}$ without reversibility; see Bernstein and Sheldon (2016) and Section 4.3.
Conditional equilibrium covariance	Fixed source $N\boldsymbol{\pi}$, its destination covariance, known N and Δ ; irreducible reversible CTMC	Recover $\mathbf{P}$ and $\mathbf{Q}$ within this class (Theorem 2).
Near-equilibrium Gaussian information	Conditional Gaussian mean and covariance derivatives; common stationary composition; nondegenerate reduced covariance	An $N^{-1/2}$ crossover in directions with nonzero mean and fluctuation sensitivities (Theorem 3).

dynamic-programming evaluation of the aggregate likelihood and a Gaussian approximation for larger populations. The latter retains both the conditional mean and covariance, making it useful in regimes where fluctuations carry much of the information. This use of latent structure to avoid brute-force enumeration is related in spirit to exact sequential computations for hidden-state models, such as the sum-product approach of Martin and Aston (2013), although the aggregation problem and recursion developed here are different.

1.1 Contributions and Main Findings

The central contribution is the conditional-covariance inversion at equilibrium. The structural-identifiability result and the Gaussian information calculation describe complementary aspects of the information retained by histogram observations; the likelihood and dynamic program provide their computational foundation.

1. **Structural identifiability and likelihood computation:** we give sufficient conditions under which known conditional histogram laws identify the labeled finite-time transition matrix. The proof uses the multiplicities of the linear factors in the conditional probability-generating functions. For statistical computation, we express the exact likelihood as a convolution of source-row multinomials, evaluate it by dynamic programming, and develop a Gaussian approximation for larger populations. These constructions use standard multinomial, convolution, and Gaussian methods tailored to the observation model.
2. **Constructive recovery from conditional equilibrium covariance:** when the source histogram is fixed at the equilibrium composition, we give an explicit inversion from the conditional destination covariance and that composition to the finite-time transition matrix and generator under reversible continuous-time dynamics. This result identifies the mechanism from a different second-order observable than the lagged cross-covariance used in stationary aggregate-moment inference.
3. **The $N^{-1/2}$ information crossover near equilibrium:** within the Gaussian aggregate approximation, if the source composition approaches equilibrium as $N^{-\alpha}$, mean-channel transition-parameter information scales as $N^{1-2\alpha}$ while fluctuation-channel information remains $O(1)$. Hence, in nondegenerate directions, mean information dominates for $\alpha < 1/2$; at $\alpha = 1/2$ the two contributions are of the same order when both are nonzero; and for $\alpha > 1/2$ the mean contribution vanishes, leaving any nonzero fluctuation information. The crossover coincides with the ordinary scale of empirical-histogram fluctuations.

The numerical studies examine the implications from several complementary directions. They compare exact and Gaussian likelihood inference across population sizes and observation lengths, examine sensitivity to temporal resolution, demonstrate constructive recovery from equilibrium fluctuations, illustrate the predicted near-equilibrium information scaling along a particular equilibrium-preserving parameter direction, and study computational and statistical performance in larger state spaces.

The remainder of the paper is organized as follows. Section 2 introduces the general aggregate CTMC observation model. Section 3 develops likelihood-based inference, dynamic-programming computation, structural identifiability, and the Gaussian approximation. Section 4 studies equilibrium and near-equilibrium information, introducing reversibility where required for the conditional-covariance inversion. Section 5 presents numerical studies based primarily on a reversible nearest-neighbor specialization, together with a matched nonlocal robustness comparison. The appendices collect technical derivations and computational details.

2 Histogram Observation Model

We consider a fixed population of $N \in \mathbb{N}$ persistent units observed at deterministic times

$$\tau_0 < \tau_1 < \cdots < \tau_T, \quad \Delta_t = \tau_t - \tau_{t-1} > 0.$$

Each unit occupies one of $K \geq 2$ states,

$$\mathcal{S} = \{1, \dots, K\}.$$

The same N units remain in the population throughout the observation period; there is no entry or exit.

For unit $i \in \{1, \dots, N\}$, let $Z_i(\tau_t) \in \mathcal{S}$ denote its hidden state at observation time τ_t . Conditional on a common generator $\mathbf{Q} = (q_{jk})$, the unit-level processes are independent and identically distributed time-homogeneous CTMCs. The generator satisfies

$$q_{jk} \geq 0 \quad (j \neq k), \quad q_{jj} = -\sum_{k \neq j} q_{jk}, \quad \mathbf{Q}\mathbf{1}_K = \mathbf{0}_K.$$

For observation interval t , the common finite-time transition matrix is

$$\mathbf{P}_t = \exp\{\Delta_t \mathbf{Q}\}, \quad P_{jk}^{(t)} = \Pr\{Z_i(\tau_t) = k \mid Z_i(\tau_{t-1}) = j\}. \quad (1)$$

Thus the Markov assumption is imposed on the persistent unit-level processes. Conditional on their current states, units transition independently according to the same matrix $\mathbf{P}_t$.

The individual states are not observed. Instead, at time τ_t we observe only the aggregate state counts

$$Y_{t,j} = \sum_{i=1}^N \mathbf{1}\{Z_i(\tau_t) = j\}, \quad j = 1, \dots, K, \quad (2)$$

and write

$$\mathbf{Y}_t = (Y_{t,1}, \dots, Y_{t,K}).$$

Hence $\sum_{j=1}^K Y_{t,j} = N$ for every t . Define the probability simplex and histogram state space by

$$\Delta_{K-1} := \{\mathbf{p} \in [0, 1]^K : \mathbf{p}\mathbf{1}_K = 1\}, \quad \mathcal{H}_{N,K} := \left\{ \mathbf{y} \in \mathbb{N}_0^K : \sum_{j=1}^K y_j = N \right\}.$$

Then $\mathbf{Y}_t \in \mathcal{H}_{N,K}$ and the empirical composition $\hat{\mathbf{p}}_t = \mathbf{Y}_t/N$ belongs to Δ_{K-1} .

Although individual transitions are hidden, the aggregate process inherits a Markov structure. Conditional on $\mathbf{Y}_{t-1} = \mathbf{y}$, the y_j units currently in source state j form a disjoint group whose destinations are distributed according to row j of $\mathbf{P}_t$. The destination histogram is therefore the sum of K independent row-wise multinomial contributions. In general it is not itself a single multinomial distribution, because the transition probabilities depend on the source state. The resulting conditional law of $\mathbf{Y}_t$ depends on the aggregate history only through $\mathbf{Y}_{t-1}$. A derivation is given in Appendix A.

3 Inference from Aggregate Histograms

Section 2 specifies how a generator $\mathbf{Q}$, through the finite-time matrices $\mathbf{P}_t = \exp\{\Delta_t \mathbf{Q}\}$, generates a histogram time series. We now address the inverse problem. The observed data are the aggregate sequence

$$\mathcal{Y}_T = (\mathbf{Y}_0, \dots, \mathbf{Y}_T),$$

whereas the unit-level transitions linking consecutive histograms are unavailable. The primary inferential object is therefore the hidden transition mechanism, represented by $\mathbf{P}_t$ and, when continuous-time recovery is considered, by $\mathbf{Q}$.

The central difficulty is that a pair of consecutive histograms determines only the row and column margins of the latent transition-count table. Many different patterns of gross individual movement can therefore produce the same observed histogram change. The observed-data likelihood must account for all transition tables compatible with those margins rather than conditioning on one unobserved flow configuration.

We first derive the exact finite-population likelihood by marginalizing over the unobserved transition counts. We then establish conditions under which the finite-time transition matrix is structurally identifiable from aggregate observations and define likelihood-based estimation. Finally, we develop a Gaussian approximation for large populations. Besides providing a scalable computational alternative, the approximation separates information carried by conditional mean changes from information carried by fluctuations, which becomes central in Section 4.

3.1 From hidden transitions to the observed-data likelihood

Consider one observation interval t . Let

$$\mathbf{y} = \mathbf{Y}_{t-1}, \quad \mathbf{y}' = \mathbf{Y}_t, \quad \mathbf{y}, \mathbf{y}' \in \mathcal{H}_{N,K},$$

and let $\mathbf{P}_t = \exp\{\Delta_t \mathbf{Q}\}$ denote the corresponding finite-time transition matrix. The likelihood for the observed pair $(\mathbf{y}, \mathbf{y}')$ is obtained by first asking what its probability would be if the individual movements, and hence the complete transition-count table, were observed.

Define the latent endpoint transition counts by

$$X_{jk}^{(t)} := \sum_{i=1}^N \mathbf{1}\{Z_i(\tau_{t-1}) = j, Z_i(\tau_t) = k\}, \quad j, k \in \mathcal{S}. \quad (3)$$

Thus $X_{jk}^{(t)}$ counts units that start the interval in state j and end it in state k , irrespective of any intermediate jumps.

For source state j , there are y_j units at the beginning of the interval. Conditional on this source count, each of those units moves independently according to row j of $\mathbf{P}_t$. Hence the j th row of the latent transition table satisfies

$$(X_{j1}^{(t)}, \dots, X_{jK}^{(t)}) \mid Y_{t-1,j} = y_j \sim \text{Multinomial}\left(y_j; P_{j1}^{(t)}, \dots, P_{jK}^{(t)}\right). \quad (4)$$

The source-state groups are disjoint, so these row vectors are conditionally independent over j . Therefore, if the complete transition table $\mathbf{X}_t = \mathbf{x} = (x_{jk})$ were observed, its conditional probability would be

$$\Pr(\mathbf{X}_t = \mathbf{x} \mid \mathbf{Y}_{t-1} = \mathbf{y}; \mathbf{P}_t) = \prod_{j=1}^K \left[\frac{y_j!}{\prod_{k=1}^K x_{jk}!} \prod_{k=1}^K P_{jk}^{(t)x_{jk}} \right]. \quad (5)$$

The transition table is not observed. The two observed histograms reveal only its row and column margins:

$$\sum_{k=1}^K x_{jk} = y_j, \quad \sum_{j=1}^K x_{jk} = y'_k.$$

Accordingly, define the set of all transition tables compatible with the observed histogram pair,

$$\mathcal{C}(\mathbf{y}, \mathbf{y}') = \left\{ \mathbf{x} \in \mathbb{N}_0^{K \times K} : \sum_{k=1}^K x_{jk} = y_j, \quad \sum_{j=1}^K x_{jk} = y'_k \right\}. \quad (6)$$

Every table in $\mathcal{C}(\mathbf{y}, \mathbf{y}')$ represents a different pattern of gross individual movement that produces the same observed net redistribution. The probability of observing $\mathbf{Y}_t = \mathbf{y}'$ given $\mathbf{Y}_{t-1} = \mathbf{y}$ is therefore obtained by summing the complete-table probabilities over all such possibilities:

$$\begin{aligned} \mathcal{K}_t(\mathbf{y}, \mathbf{y}'; \mathbf{P}_t) &:= \Pr(\mathbf{Y}_t = \mathbf{y}' \mid \mathbf{Y}_{t-1} = \mathbf{y}; \mathbf{P}_t) \\ &= \sum_{\mathbf{x} \in \mathcal{C}(\mathbf{y}, \mathbf{y}')} \Pr(\mathbf{X}_t = \mathbf{x} \mid \mathbf{Y}_{t-1} = \mathbf{y}; \mathbf{P}_t) \\ &= \sum_{\mathbf{x} \in \mathcal{C}(\mathbf{y}, \mathbf{y}')} \prod_{j=1}^K \left[\frac{y_j!}{\prod_{k=1}^K x_{jk}!} \prod_{k=1}^K P_{jk}^{(t)x_{jk}} \right]. \end{aligned} \quad (7)$$

Thus the exact aggregate likelihood explicitly accounts for the fact that the observed histogram change determines the margins of the hidden flow table, but not the table itself.

The same construction shows that the aggregate histogram process is itself Markov under the present independent common unit-level mechanism: conditional on the current histogram, the destination counts depend on no earlier aggregate history. A derivation is provided in Appendix A. Consequently, conditioning on the initial histogram $\mathbf{Y}_0$, the likelihood of the observed sequence factorizes as

$$L_T(\mathbf{Q} \mid \mathbf{Y}_0) = \prod_{t=1}^T \mathcal{K}_t(\mathbf{Y}_{t-1}, \mathbf{Y}_t; \mathbf{P}_t(\mathbf{Q})), \quad \ell_T(\mathbf{Q} \mid \mathbf{Y}_0) = \sum_{t=1}^T \log \mathcal{K}_t(\mathbf{Y}_{t-1}, \mathbf{Y}_t; \mathbf{P}_t(\mathbf{Q})). \quad (8)$$

No probability model for the initial histogram is required for this conditional-likelihood analysis. In the remainder of the section we suppress the explicit conditioning on $\mathbf{Y}_0$ when no ambiguity can arise and write $L_T(\mathbf{Q})$ and $\ell_T(\mathbf{Q})$ for brevity.

3.2 Structural identifiability

A realized latent transition table is generally not recoverable from its row and column margins. This does not imply that the persistent mechanism is structurally unidentified. Identifiability concerns the map from the hidden transition mechanism to the family of observable conditional histogram laws, not reconstruction of any particular hidden transition table. The first question is therefore whether the labeled finite-time transition matrix $\mathbf{P}$ is determined by those observable laws.

For a fixed observation interval Δ , write $\mathbf{P} = \exp\{\Delta\mathbf{Q}\}$ and define the conditional multivariate probability-generating function of the destination histogram by

$$G_{\mathbf{y}}(\mathbf{s}) := \mathbb{E} \left[\prod_{k=1}^K s_k^{Y_{t,k}} \middle| \mathbf{Y}_{t-1} = \mathbf{y} \right], \quad \mathbf{s} = (s_1, \dots, s_K). \quad (9)$$

Since the y_j units originating in state j move independently according to row j of $\mathbf{P}$, conditional independence across source states gives

$$G_{\mathbf{y}}(\mathbf{s}) = \prod_{j=1}^K \left(\sum_{k=1}^K P_{jk} s_k \right)^{y_j}. \quad (10)$$

Thus the observable aggregate-transition law is encoded by a product of linear factors, with the j th factor determined by the j th row of $\mathbf{P}$.

Equation (10) first gives a simple identifiability condition from a single source histogram. If every source count y_j is positive and the counts $y_1, \dots, y_K$ are pairwise distinct, then the multiplicity of each recovered linear factor identifies the source state to which that factor belongs. Since the factor itself is determined by the corresponding row of $\mathbf{P}$, the labeled transition matrix is then identifiable from the conditional aggregate-transition law.

Requiring all counts in one source histogram to be distinct is sufficient but restrictive. The same labeling information can instead be accumulated across several source compositions. To express this generalization, consider M source histograms

$$\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(M)} \in \mathcal{H}_{N,K}$$

for which the corresponding conditional aggregate-transition laws are known. Here M is not an additional parameter of the observation model and does not represent parallel Markov chains; it simply indexes the source compositions used in the identifiability argument. For source state j , the counts across these compositions form the source-count profile

$$\mathbf{v}_j := (y_j^{(1)}, \dots, y_j^{(M)})^\top.$$

If these profiles are nonzero and pairwise distinct, they provide unique labels for the linear factors associated with the rows of $\mathbf{P}$. This gives the following sufficient condition. Appendix B provides the factorization argument. No realized latent transition table needs to be recovered.

Theorem 1 (Structural identifiability from aggregate histograms).

Consider the general CTMC observation model of Section 2 over a fixed observation interval. Suppose that the conditional aggregate-transition laws are known for source histograms $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(M)}$. If every source-count profile

$$\mathbf{v}_j = (y_j^{(1)}, \dots, y_j^{(M)})^\top$$

is nonzero and the profiles are pairwise distinct, then the labeled finite-time transition matrix $\mathbf{P}$ is uniquely determined by these conditional aggregate-transition laws.

Proof. See Appendix B. □

As an immediate single-source special case, if $M = 1$, the labeled matrix $\mathbf{P}$ is identifiable whenever all K source counts are positive and pairwise distinct.

Thus the labeled finite-time transition matrix can be identified without observing or reconstructing the latent transition tables. This result does not require reversibility. Recovering a continuous-time generator $\mathbf{Q}$ from an identified $\mathbf{P}$ is a separate embedding problem and may require additional assumptions; the reversible setting of Theorem 2 provides one such case.

The theorem is deliberately a population-level structural statement. It assumes the relevant conditional laws are known and asks whether they determine a unique finite-time mechanism. A finite realized trajectory poses the separate statistical question of how accurately that mechanism can be learned from the observed source histograms.

3.3 Maximum-likelihood estimation

Theorem 1 is a population-level identifiability result: it asks whether known conditional aggregate-transition laws determine the finite-time transition matrix. In practice, these laws are not known; instead, we observe a finite histogram sequence $\mathbf{Y}_0, \dots, \mathbf{Y}_T$. We therefore estimate the transition dynamics by maximum likelihood under the general CTMC observation model of Section 2. Let

$$\mathcal{Q} = \{\mathbf{Q} \in \mathbb{R}^{K \times K} : q_{jk} \geq 0 \text{ for } j \neq k, \quad \mathbf{Q}\mathbf{1}_K = \mathbf{0}_K\} \quad (11)$$

denote the space of valid generators. For each $\mathbf{Q} \in \mathcal{Q}$ and observation interval Δ_t , the corresponding finite-time transition matrix is

$$\mathbf{P}_t(\mathbf{Q}) = \exp\{\Delta_t \mathbf{Q}\}.$$

Thus the conditional sequence likelihood is

$$L_T(\mathbf{Q} \mid \mathbf{Y}_0) = \prod_{t=1}^T \mathcal{K}_t(\mathbf{Y}_{t-1}, \mathbf{Y}_t; \mathbf{P}_t(\mathbf{Q})),$$

with log-likelihood $\ell_T(\mathbf{Q} \mid \mathbf{Y}_0)$. As above, we suppress the explicit conditioning on $\mathbf{Y}_0$ when no ambiguity can arise.

For a finite histogram sequence, an exact maximum-likelihood estimator is any element of

$$\hat{\mathbf{Q}} \in \arg \max_{\mathbf{Q} \in \mathcal{Q}} \ell_T(\mathbf{Q}), \quad (12)$$

provided this set is nonempty. The likelihood conditions on the observed initial histogram $\mathbf{Y}_0$; hence no separate probability model for the initial population composition is required.

The generator space $\mathcal{Q}$ is noncompact, and existence of an unrestricted maximizer is therefore not automatic. Structural identifiability of the finite-time transition matrix does not by itself resolve this finite-sample optimization issue. In numerical work one may instead optimize over a prespecified compact subset of $\mathcal{Q}$; any such restriction should be reported as part of the estimation procedure. Appendix C gives explicit examples of boundary behavior showing why an unrestricted maximizer need not exist.

The contingency-table sum in (7) is the observed-data likelihood: the latent transition counts are not estimated as an intermediate step, but are marginalized over all transition tables compatible with the observed source and destination histograms. Direct evaluation of this sum by enumerating all feasible tables is unnecessary. Conditional on the source histogram, each source state contributes an independent multinomial destination-count vector, and the observed destination histogram is their sum. The exact transition probability can therefore be evaluated as a finite convolution using a forward dynamic program. A companion backward recursion yields conditional transition-count moments and hence derivatives of the exact likelihood without enumerating the compatible tables. Detailed forward-backward recursions are given in Appendix D.

For numerical optimization, let $\boldsymbol{\vartheta}$ denote an unconstrained parameter vector mapped to an admissible generator $\mathbf{Q}(\boldsymbol{\vartheta})$. The particular reversible nearest-neighbor parameterization used in the numerical studies is introduced in Section 5.

Algorithm 1 summarizes the numerical fitting procedure for the exact MLE. Each conditional transition probability is evaluated using the forward dynamic program of Appendix D, with the companion backward recursion providing the gradient when used, and the resulting likelihood in (8) is optimized over $\boldsymbol{\vartheta}$. Multiple initializations reduce sensitivity to local numerical behavior but, because global concavity is not assumed, do not guarantee a global optimum.

The computational burden of Algorithm 1 has two distinct components. For a common observation interval Δ , computing $\mathbf{P} = \exp\{\Delta \mathbf{Q}\}$ with dense methods costs $O(K^3)$ per likelihood

Algorithm 1 Numerical maximum-likelihood fitting from histogram time series

Require: Histograms $\mathbf{Y}_0, \dots, \mathbf{Y}_T$, intervals $\Delta_1, \dots, \Delta_T$, and R_{init} initial values $\{\boldsymbol{\vartheta}_0^{(r)}\}_{r=1}^{R_{\text{init}}}$.

- 1: **for** $r = 1, \dots, R_{\text{init}}$ **do**
- 2: Initialize $\boldsymbol{\vartheta} \leftarrow \boldsymbol{\vartheta}_0^{(r)}$.
- 3: **repeat**
- 4: Map $\boldsymbol{\vartheta} \mapsto \mathbf{Q}(\boldsymbol{\vartheta})$ and construct the distinct $\mathbf{P}_t = \exp\{\Delta_t \mathbf{Q}(\boldsymbol{\vartheta})\}$.
- 5: Evaluate each aggregate transition probability by the forward dynamic program in Appendix D.
- 6: Accumulate the exact log-likelihood and, when used, its gradient from the forward-backward recursion.
- 7: Apply a line-search quasi-Newton update.
- 8: **until** the convergence criteria are satisfied
- 9: Store the converged parameter and attained likelihood.
- 10: **end for**
- 11: Retain the converged solution with the largest attained likelihood.

evaluation, while direct Fréchet differentiation with respect to d_ϑ parameters costs up to $O(d_\vartheta K^3)$. The remaining cost comes from the exact aggregate forward-backward recursion. Let $C_{\text{DP},t}$ denote the unpruned cost of evaluating the transition probability and posterior flow moments for observed transition t ; Appendix D derives this quantity explicitly in terms of the forward and backward message supports and the possible source-row compositions.

If initialization r requires I_r likelihood/gradient evaluations, an upper-bound description of the computational load is therefore

$$C_{\text{Alg 1}} = O\left\{\sum_{r=1}^{R_{\text{init}}} I_r \left[(1 + d_\vartheta) K^3 + \sum_{t=1}^T C_{\text{DP},t} \right]\right\}. \quad (13)$$

This expression separates matrix-exponential differentiation from the exact aggregate recursion, while also making explicit the effects of observation length, optimization iterations, and multiple initializations. Appendix D shows that, for fixed K , the message-state counts are polynomial in N , but their degree increases with K ; when both N and K grow, the exact recursion rapidly becomes expensive. This population-dependent dynamic-programming bottleneck motivates the large-population approximation considered next.

3.4 Gaussian Approximation for Large Populations

For large populations, repeatedly evaluating the exact finite-population likelihood can become expensive. We therefore use a Gaussian approximation to the conditional aggregate-transition law. The approximation does not change the data-generating model or the underlying generator $\mathbf{Q}$; it replaces only the exact conditional probability mass used in likelihood evaluation. Besides its computational role, the Gaussian form makes explicit the two ways in which the transition mechanism affects the observations: through the conditional mean and through the conditional covariance. This separation is central to the equilibrium analysis in Section 4.

Conditionally on the observed source histogram $\mathbf{Y}_{t-1}$, the destination histogram is the sum of independent categorical contributions from the N persistent units. Define the empirical composition

$$\hat{\mathbf{p}}_t = \frac{\mathbf{Y}_t}{N} \in \mathcal{P}_N, \quad \mathcal{P}_N := \{\mathbf{p} \in [0, 1]^K : N\mathbf{p} \in \mathbb{N}_0^K, \mathbf{p}\mathbf{1}_K = 1\}.$$

Because the K histogram components sum to N , one coordinate is redundant and the full covariance matrix is singular. We therefore set $d = K - 1$ and, for any K -vector $\mathbf{a}$, write $\mathbf{a}_- \in \mathbb{R}^d$

for the column vector $(a_1, \dots, a_d)^\top$ obtained by deleting its final component, irrespective of whether $\mathbf{a}$ is originally written as a row or column vector. Full histogram compositions remain row vectors.

For source state j , let

$$\mathbf{p}_{t,j,-} := (P_{j1}^{(t)}, \dots, P_{j,K-1}^{(t)})^\top$$

denote the first $K - 1$ destination probabilities in row j of $\mathbf{P}_t(\mathbf{Q})$. Conditional on the observed source histogram, the $Y_{t-1,j}$ individuals in source state j contribute a multinomial destination-count vector with probabilities given by row j of $\mathbf{P}_t$. The destination histogram is the sum of these independent source-state contributions. Consequently, its conditional mean and covariance are obtained by summing the corresponding multinomial moments. The exact conditional mean composition in reduced coordinates is

$$\mathbf{m}_t(\mathbf{Q}) = (\hat{\mathbf{p}}_{t-1} \mathbf{P}_t(\mathbf{Q}))_-, \quad (14)$$

and the exact conditional covariance per unit is

$$\mathbf{V}_t(\mathbf{Q}) = \sum_{j=1}^K \hat{p}_{t-1,j} \left[\text{diag}(\mathbf{p}_{t,j,-}) - \mathbf{p}_{t,j,-} \mathbf{p}_{t,j,-}^\top \right]. \quad (15)$$

Consequently,

$$\mathbb{E}[\mathbf{Y}_{t,-} \mid \mathbf{Y}_{t-1}] = N \mathbf{m}_t(\mathbf{Q}), \quad \text{Cov}(\mathbf{Y}_{t,-} \mid \mathbf{Y}_{t-1}) = N \mathbf{V}_t(\mathbf{Q}). \quad (16)$$

These moments are properties of the exact finite-population model; the approximation enters only in replacing the discrete conditional law by a Gaussian law with the same first two moments:

$$\mathbf{Y}_{t,-} \mid \mathbf{Y}_{t-1} \approx \mathcal{N}_d(N \mathbf{m}_t(\mathbf{Q}), N \mathbf{V}_t(\mathbf{Q})). \quad (17)$$

For fixed K and a nondegenerate reduced conditional covariance, the standardized aggregate count vector converges to the corresponding multivariate Gaussian law as N grows. Using this approximation for individual lattice probabilities requires additional local-limit regularity conditions. Appendix E gives the distributional central-limit argument, while Appendix F states the conditions for the local likelihood approximation.

Define the reduced composition residual

$$\mathbf{r}_t(\mathbf{Q}) = \hat{\mathbf{p}}_{t,-} - \mathbf{m}_t(\mathbf{Q}) = \hat{\mathbf{p}}_{t,-} - (\hat{\mathbf{p}}_{t-1} \mathbf{P}_t(\mathbf{Q}))_-. \quad (18)$$

Substituting the Gaussian transition law in (17) for each observation interval gives the approximate conditional sequence log-likelihood

$$\tilde{\ell}_T(\mathbf{Q}) = -\frac{1}{2} \sum_{t=1}^T \left[\log \det \mathbf{V}_t(\mathbf{Q}) + N \mathbf{r}_t(\mathbf{Q})^\top \mathbf{V}_t(\mathbf{Q})^{-1} \mathbf{r}_t(\mathbf{Q}) \right] + C_N, \quad (19)$$

where

$$C_N = -\frac{T(K-1)}{2} \log(2\pi N) \quad (20)$$

does not depend on $\mathbf{Q}$. It can therefore be omitted during optimization.

The Gaussian criterion is defined only when every reduced conditional covariance is positive definite. For the observed sequence, define the data-dependent domain

$$\mathcal{Q}_G(\mathcal{Y}_T) := \{\mathbf{Q} \in \mathcal{Q} : \mathbf{V}_t(\mathbf{Q}) \succ 0 \text{ for } t = 1, \dots, T\}.$$

In particular, $\mathbf{Q} = \mathbf{0}$ is excluded. When the Gaussian criterion attains its maximum on this domain, define the Gaussian approximate maximum-likelihood estimator by

$$\hat{\mathbf{Q}}_G \in \arg \max_{\mathbf{Q} \in \mathcal{Q}_G(\mathcal{Y}_T)} \tilde{\ell}_T(\mathbf{Q}). \quad (21)$$

The same optimization framework as Algorithm 1 can be used for the Gaussian estimator, replacing the exact conditional transition probability and its score by the Gaussian criterion in (19) and its gradient. Thus the two estimators differ in the conditional likelihood evaluation, while the parameterization, multiple-initialization strategy, and numerical optimization framework can be kept the same. Algorithm 1 describes a possible gradient-based implementation; some numerical comparisons below instead use derivative-free optimization, as specified in Appendix J. The exact and Gaussian criteria therefore target the same general CTMC generator; they differ only in how the conditional histogram-transition probability is evaluated. Existence of a Gaussian maximizer on $\mathcal{Q}_G(\mathcal{Y}_T)$ is not automatic. On a nonempty compact subset of this domain, the covariance eigenvalues are uniformly bounded away from zero, the criterion is continuous, and a maximizer exists. Any such numerical restriction should be reported explicitly; compactness within $\mathcal{Q}$ alone does not exclude singular covariances.

The computational benefit can be stated directly relative to Algorithm 1. Once $\mathbf{P}_t$, $\mathbf{m}_t$, and $\mathbf{V}_t$ have been formed, the Gaussian criterion uses a $(K - 1)$ -dimensional covariance matrix and has essentially no dependence on the population size N . With dense linear algebra, covariance factorization and the associated solves cost $O(K^3)$ per observed transition. Thus, for a common observation interval, the Gaussian criterion replaces the population-dependent dynamic-programming term $C_{DP,t}$ in (13) by dense operations on K -dimensional quantities. A dense implementation has per-evaluation likelihood cost $O(K^3 + TK^3)$ after lower-order moment construction, with additional parameter-derivative costs when an analytic gradient is used. The optimization-iteration and multiple-initialization factors remain. The essential computational gain is therefore removal of the exact population-dependent dynamic program, not removal of dependence on the state-space dimension.

More importantly for the theory that follows, the transition mechanism enters the Gaussian likelihood through two distinct objects: the conditional mean $\mathbf{m}_t$ and the conditional covariance $\mathbf{V}_t$. Parameter sensitivity of the mean provides a mean-redistribution channel, while parameter sensitivity of the covariance provides a fluctuation channel through both the Mahalanobis weighting and the log-determinant term. For transition families that preserve a common stationary composition, the mean channel vanishes at an exact equilibrium source, whereas the covariance channel can remain informative. Section 4 develops this distinction.

4 Information About Transition Dynamics at and Near Equilibrium

Sections 2-3 considered a general continuous-time Markov chain and did not require reversibility. We now study what can be learned about the transition dynamics when the observed population composition is at, or close to, equilibrium.

For this purpose, assume that the generator $\mathbf{Q}$ is irreducible, and let $\boldsymbol{\pi} \in \text{int}(\Delta_{K-1})$ denote its unique stationary distribution,

$$\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}_K^\top.$$

For a fixed observation interval of length $\Delta > 0$, write

$$\mathbf{P} = \exp\{\Delta\mathbf{Q}\}, \quad \boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}, \quad \mathbf{D}_\pi = \text{diag}(\boldsymbol{\pi}).$$

We first derive the conditional moments at an exact equilibrium source using only stationarity. Reversibility is imposed later, for the inversion result in Theorem 2. As in Section 3, covariance matrices are represented in the first $K - 1$ coordinates when an inverse is required.

4.1 Conditional Moments at Equilibrium

For a source histogram $\mathbf{y}$, the conditional moments of the row-wise multinomial transition counts give

$$\mathbb{E}[\mathbf{Y}_t \mid \mathbf{Y}_{t-1} = \mathbf{y}] = \mathbf{y}\mathbf{P}, \quad (22)$$

and, writing $\mathbf{p}_j$ for row j of $\mathbf{P}$ as a row vector,

$$\text{Cov}(\mathbf{Y}_t \mid \mathbf{Y}_{t-1} = \mathbf{y}) = \sum_{j=1}^K y_j \left[\text{diag}(\mathbf{p}_j) - \mathbf{p}_j^\top \mathbf{p}_j \right]. \quad (23)$$

These identities follow directly from the aggregate transition model and do not themselves require stationarity.

Here, *exact aggregate equilibrium* means that the observed source composition equals the stationary composition, $\mathbf{Y}_{t-1}/N = \boldsymbol{\pi}$. This is a conditional statement about the observed histogram; it does not assert that a finite population sampled under stationarity has a nonrandom histogram.

Assume $N\pi_j \in \mathbb{N}_0$ for all j and condition on $\mathbf{Y}_{t-1} = N\boldsymbol{\pi}$. Since $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$, (22) gives

$$\mathbb{E}[\mathbf{Y}_t \mid \mathbf{Y}_{t-1} = N\boldsymbol{\pi}] = N\boldsymbol{\pi}. \quad (24)$$

Thus the conditional mean is identical for all transition mechanisms having stationary distribution $\boldsymbol{\pi}$.

In contrast, substituting $y_j = N\pi_j$ into (23) gives the general equilibrium conditional covariance

$$\boldsymbol{\Sigma}_\pi := \text{Cov}(\mathbf{Y}_t \mid \mathbf{Y}_{t-1} = N\boldsymbol{\pi}) = N \left(\mathbf{D}_\pi - \mathbf{P}^\top \mathbf{D}_\pi \mathbf{P} \right). \quad (25)$$

The covariance therefore generally retains information about the transition mechanism even when the conditional mean does not.

If, in addition, the CTMC is reversible with respect to $\boldsymbol{\pi}$, then

$$\mathbf{P}^\top \mathbf{D}_\pi = \mathbf{D}_\pi \mathbf{P},$$

and (25) reduces to

$$\boldsymbol{\Sigma}_\pi = N \mathbf{D}_\pi (\mathbf{I}_K - \mathbf{P}^2). \quad (26)$$

This additional reversible structure makes the explicit inversion in the next subsection possible.

4.2 Recovery of Transition Dynamics from Equilibrium Conditional Covariance

Under reversibility, define the similarity transform

$$\mathbf{S} = \mathbf{D}_\pi^{1/2} \mathbf{P} \mathbf{D}_\pi^{-1/2} \quad (27)$$

and the standardized equilibrium covariance

$$\mathbf{C}_\pi = \frac{1}{N} \mathbf{D}_\pi^{-1/2} \boldsymbol{\Sigma}_\pi \mathbf{D}_\pi^{-1/2}. \quad (28)$$

Equation (26) then gives

$$\mathbf{C}_\pi = \mathbf{I}_K - \mathbf{S}^2. \quad (29)$$

Thus $(\boldsymbol{\pi}, \boldsymbol{\Sigma}_\pi)$ determines $\mathbf{S}^2$. In general this would leave a square-root ambiguity. Reversibility makes $\mathbf{S}$ symmetric, while the continuous-time representation resolves the remaining ambiguity: with

$$\mathbf{H} = \mathbf{D}_\pi^{1/2} \mathbf{Q} \mathbf{D}_\pi^{-1/2},$$

reversibility makes $\mathbf{H}$ real symmetric and

$$\mathbf{S} = e^{\Delta \mathbf{H}} \succ 0.$$

Hence $\mathbf{S}$ is the unique symmetric positive-definite principal square root of $\mathbf{I}_K - \mathbf{C}_\pi$. This yields the following recovery result.

Theorem 2 (Recovery from equilibrium conditional covariance).

Suppose that $\mathbf{Q}$ is the generator of an irreducible CTMC with stationary distribution $\boldsymbol{\pi}$, and condition on $\mathbf{Y}_{t-1} = N\boldsymbol{\pi}$, where $N\pi_j \in \mathbb{N}_0$ for all j . Let

$$\boldsymbol{\Sigma}_\pi = \text{Cov}(\mathbf{Y}_t \mid \mathbf{Y}_{t-1} = N\boldsymbol{\pi})$$

for a known observation interval $\Delta > 0$.

If, in addition, the CTMC is reversible with respect to $\boldsymbol{\pi}$, then the pair $(\boldsymbol{\pi}, \boldsymbol{\Sigma}_\pi)$ uniquely determines the finite-time transition matrix $\mathbf{P} = \exp\{\Delta \mathbf{Q}\}$ and the generator $\mathbf{Q}$.

Proof. See Appendix G. □

The recovery is constructive:

$$(\boldsymbol{\pi}, \boldsymbol{\Sigma}_\pi) \longrightarrow \mathbf{C}_\pi \longrightarrow \mathbf{S} = (\mathbf{I}_K - \mathbf{C}_\pi)^{1/2} \longrightarrow \mathbf{P} \longrightarrow \mathbf{Q}. \quad (30)$$

Specifically,

$$\mathbf{P} = \mathbf{D}_\pi^{-1/2} (\mathbf{I}_K - \mathbf{C}_\pi)^{1/2} \mathbf{D}_\pi^{1/2}, \quad (31)$$

and

$$\mathbf{Q} = \frac{1}{2\Delta} \mathbf{D}_\pi^{-1/2} \log(\mathbf{I}_K - \mathbf{C}_\pi) \mathbf{D}_\pi^{1/2}. \quad (32)$$

Reversibility and continuous time play distinct roles in this inversion: reversibility produces the symmetric similarity transform, while continuous-time embeddability makes that transform positive definite and therefore selects the principal square root. These facts are closely related to the reversible embedding problem; see, e.g., (Jia, 2016).

Theorem 2 is a population-level identifiability result for the exact conditional second-order law. When the source histogram is $N\boldsymbol{\pi}$, the equilibrium composition is observed directly as $\boldsymbol{\pi} = \mathbf{Y}_{t-1}/N$; the additional quantity required for inversion is the conditional covariance $\boldsymbol{\Sigma}_\pi$. The theorem does not imply that $\boldsymbol{\Sigma}_\pi$ can be estimated accurately from a single endpoint pair.

Direct empirical use of this inversion requires an estimate of the conditional covariance at $N\boldsymbol{\pi}$. Replicated transitions from that fixed source provide one design, as illustrated in Experiment 2. Pooling transitions from nearby source compositions introduces an additional approximation whose bias must be controlled. A single stationary trajectory visits random source histograms around $N\boldsymbol{\pi}$, so the theorem alone does not provide a covariance estimator for that observation design. Exact conditioning also requires $N\pi_j$ to be integral for every state, as assumed in the theorem.

The inversion is exact for population covariances generated by the assumed model. With an estimated covariance, a positive-definite matrix inside the square root and logarithm does not by itself guarantee nonnegative entries of the recovered $\mathbf{P}$ or nonnegative off-diagonal entries of the recovered $\mathbf{Q}$. Finite-sample feasibility and regularization are therefore separate estimation questions.

4.3 Relation to Stationary Aggregate Moments

The equilibrium conditional covariance used in Theorem 2 should be distinguished from the lagged covariance of a population observed under its stationary distribution. Suppose the

independent units are initialized in stationarity, so that the source histogram is random. Define the lag-one aggregate cross-covariance, using the row-vector convention, by

$$\mathbf{\Gamma}_1 := \mathbb{E} \left[(\mathbf{Y}_{t-1} - N\boldsymbol{\pi})^\top (\mathbf{Y}_t - N\boldsymbol{\pi}) \right].$$

Independence across units and the joint state probabilities $\pi_j P_{jk}$ give

$$\mathbf{\Gamma}_1 = N \left(\mathbf{D}_\pi \mathbf{P} - \boldsymbol{\pi}^\top \boldsymbol{\pi} \right), \quad \mathbf{P} = \mathbf{D}_\pi^{-1} \left(\frac{\mathbf{\Gamma}_1}{N} + \boldsymbol{\pi}^\top \boldsymbol{\pi} \right).$$

Thus, when $\boldsymbol{\pi}$ is known, stationary lag-one covariance identifies $\mathbf{P}$ without requiring reversibility. This is the stationary specialization of the aggregate-moment identities used by Bernstein and Sheldon (2016).

Theorem 2 uses a different second-order observable. There the source histogram is conditioned to be exactly $N\boldsymbol{\pi}$, so its randomness is removed and no cross-time covariance is available. Identification instead comes from

$$\boldsymbol{\Sigma}_\pi = \text{Cov}(\mathbf{Y}_t \mid \mathbf{Y}_{t-1} = N\boldsymbol{\pi}),$$

the covariance of the destination histogram conditional on a fixed equilibrium source. Under reversible continuous-time dynamics, this conditional covariance determines $\mathbf{P}$ through the square-root inversion above and then determines $\mathbf{Q}$.

Stationary lag covariance and equilibrium conditional covariance therefore provide distinct routes to transition information. The former exploits temporal dependence between two random aggregate histograms; the latter exploits endpoint fluctuations conditional on a fixed source composition. The contribution of Theorem 2 is the latter inversion, rather than a general claim that stationary aggregate moments identify transition dynamics.

4.4 Mean-Fluctuation Information Regimes Near Equilibrium

The preceding results concern population-level structural recovery from conditional moments. We now use the Gaussian approximation to quantify how statistical sensitivity shifts between conditional-mean and conditional-covariance information as a finite observed source composition approaches equilibrium.

The Fisher information below is computed under the Gaussian conditional model itself. The local probability approximation in Appendix F does not justify differentiating its remainder or transferring its information to the exact histogram likelihood. Such a transfer would require additional control of likelihood derivatives. The result also holds the stationary composition fixed across the parameter family; joint estimation of composition and transition rates involves further information terms.

Let

$$\boldsymbol{\rho} = \mathbf{Y}_{t-1}/N$$

denote the empirical source composition. Throughout this subsection we condition on $\mathbf{Y}_{t-1} = N\boldsymbol{\rho}$, so $\boldsymbol{\rho}$ is treated as fixed when computing conditional Fisher information. We work in reduced coordinates of dimension $d = K - 1$.

The result concerns parameter changes that alter the transition dynamics while preserving the equilibrium composition. Let $\boldsymbol{\eta} = (\eta_1, \dots, \eta_m)$ parameterize a continuously differentiable family of transition matrices $\mathbf{P}(\boldsymbol{\eta})$ having the common stationary distribution $\boldsymbol{\pi}$:

$$\boldsymbol{\pi} \mathbf{P}(\boldsymbol{\eta}) = \boldsymbol{\pi} \quad \text{for every admissible } \boldsymbol{\eta}. \quad (33)$$

This common-stationary-distribution assumption is essential to the disappearance of the mean channel at exact equilibrium. No reversibility assumption is required in this subsection.

For source state j , let

$$\mathbf{p}_{j,-}(\boldsymbol{\eta}) := (P_{j1}(\boldsymbol{\eta}), \dots, P_{jd}(\boldsymbol{\eta}))^\top$$

and define

$$\mathbf{C}_j(\boldsymbol{\eta}) = \text{diag}(\mathbf{p}_{j,-}(\boldsymbol{\eta})) - \mathbf{p}_{j,-}(\boldsymbol{\eta})\mathbf{p}_{j,-}(\boldsymbol{\eta})^\top, \quad \mathbf{V}_\rho(\boldsymbol{\eta}) = \sum_{j=1}^K \rho_j \mathbf{C}_j(\boldsymbol{\eta}). \quad (34)$$

The conditional Gaussian approximation of Section 3 can then be written as

$$\mathbf{Y}_{t,-} \mid \mathbf{Y}_{t-1} = N\boldsymbol{\rho} \approx \mathcal{N}_d(N\mathbf{m}_\rho(\boldsymbol{\eta}), N\mathbf{V}_\rho(\boldsymbol{\eta})), \quad \mathbf{m}_\rho(\boldsymbol{\eta}) = (\boldsymbol{\rho}\mathbf{P}(\boldsymbol{\eta}))_-. \quad (35)$$

For $r = 1, \dots, m$, write

$$\mathbf{m}_r(\boldsymbol{\rho}, \boldsymbol{\eta}) = \frac{\partial \mathbf{m}_\rho(\boldsymbol{\eta})}{\partial \eta_r}, \quad \mathbf{V}_r(\boldsymbol{\rho}, \boldsymbol{\eta}) = \frac{\partial \mathbf{V}_\rho(\boldsymbol{\eta})}{\partial \eta_r},$$

and

$$\mathbf{P}_r = \frac{\partial \mathbf{P}(\boldsymbol{\eta})}{\partial \eta_r}.$$

For $r, s = 1, \dots, m$, the standard Fisher-information formula for a multivariate Gaussian model with parameter-dependent mean and covariance gives (Kay, 1993)

$$\mathcal{I}_{rs}^{(N)}(\boldsymbol{\rho}, \boldsymbol{\eta}) = \underbrace{N\mathbf{m}_r^\top \mathbf{V}_\rho^{-1} \mathbf{m}_s}_{\text{mean channel}} + \underbrace{\frac{1}{2} \text{tr}(\mathbf{V}_\rho^{-1} \mathbf{V}_r \mathbf{V}_\rho^{-1} \mathbf{V}_s)}_{\text{fluctuation channel}}. \quad (36)$$

Differentiating (33) with respect to η_r gives

$$\boldsymbol{\pi} \mathbf{P}_r = \mathbf{0}_K^\top.$$

Therefore

$$\mathbf{m}_r(\boldsymbol{\rho}, \boldsymbol{\eta}) = [(\boldsymbol{\rho} - \boldsymbol{\pi})\mathbf{P}_r]_-. \quad (37)$$

This identity is the mechanism behind the information transition. At exact equilibrium, $\boldsymbol{\rho} = \boldsymbol{\pi}$ implies $\mathbf{m}_r = \mathbf{0}$, so parameter sensitivity through the conditional mean vanishes for the equilibrium-preserving family. The conditional covariance can nevertheless remain parameter-sensitive, leaving a fluctuation-information channel.

Suppose now that the source composition approaches equilibrium at rate

$$\|\boldsymbol{\rho}_N - \boldsymbol{\pi}\| = O(N^{-\alpha}).$$

Then (37) gives $\|\mathbf{m}_r(\boldsymbol{\rho}_N, \boldsymbol{\eta})\| = O(N^{-\alpha})$. The mean contribution to the Gaussian Fisher information is therefore $O(N^{1-2\alpha})$, whereas the fluctuation contribution remains $O(1)$ under the regularity conditions below. Consequently, $N^{-1/2}$ is the scale at which the two channels can contribute at the same order; whether either channel is present in a particular parameter direction depends on the corresponding sensitivity.

Theorem 3 (Mean-fluctuation information regimes near equilibrium).

Let $\boldsymbol{\eta} \in \mathbb{R}^m$ parameterize a family of transition matrices $\mathbf{P}(\boldsymbol{\eta})$ such that

$$\boldsymbol{\pi} \mathbf{P}(\boldsymbol{\eta}) = \boldsymbol{\pi}$$

for every $\boldsymbol{\eta}$ in a neighborhood of the true parameter. Assume that $\mathbf{P}(\boldsymbol{\eta})$ is continuously differentiable in $\boldsymbol{\eta}$, that $\mathbf{V}_\rho(\boldsymbol{\eta})$ and its first derivatives with respect to $\boldsymbol{\eta}$ are jointly continuous in $(\boldsymbol{\rho}, \boldsymbol{\eta})$ in a neighborhood of $(\boldsymbol{\pi}, \boldsymbol{\eta})$, and that $\mathbf{V}_\pi(\boldsymbol{\eta})$ is positive definite in reduced coordinates.

Consider a sequence of admissible empirical source compositions $\boldsymbol{\rho}_N$, with $N\boldsymbol{\rho}_N \in \mathbb{N}_0^K$, satisfying

$$\boldsymbol{\rho}_N = \boldsymbol{\pi} + N^{-\alpha}\mathbf{h} + o(N^{-\alpha}), \quad \alpha > 0, \quad \mathbf{h}\mathbf{1}_K = 0, \quad \mathbf{h} \in \mathbb{R}^K. \quad (38)$$

Then the Fisher information in the Gaussian aggregate-transition approximation satisfies

$$\mathcal{I}_{rs}^{(N)}(\boldsymbol{\rho}_N, \boldsymbol{\eta}) = N^{1-2\alpha} A_{rs}(\mathbf{h}) + B_{rs} + o(N^{1-2\alpha}) + o(1), \quad (39)$$

where

$$\mathbf{V}_{\boldsymbol{\pi},r} := \left. \frac{\partial \mathbf{V}_{\boldsymbol{\rho}}(\boldsymbol{\eta})}{\partial \eta_r} \right|_{\boldsymbol{\rho}=\boldsymbol{\pi}},$$

and

$$A_{rs}(\mathbf{h}) = [\mathbf{h}\mathbf{P}_r]_-^\top \mathbf{V}_{\boldsymbol{\pi}}^{-1} [\mathbf{h}\mathbf{P}_s]_-, \quad (40)$$

$$B_{rs} = \frac{1}{2} \text{tr}(\mathbf{V}_{\boldsymbol{\pi}}^{-1} \mathbf{V}_{\boldsymbol{\pi},r} \mathbf{V}_{\boldsymbol{\pi}}^{-1} \mathbf{V}_{\boldsymbol{\pi},s}). \quad (41)$$

Write

$$A(\mathbf{h}) := (A_{rs}(\mathbf{h}))_{r,s=1}^m, \quad B := (B_{rs})_{r,s=1}^m.$$

For any fixed parameter-space direction $\mathbf{a} \in \mathbb{R}^m$,

$$\mathbf{a}^\top \mathcal{I}^{(N)}(\boldsymbol{\rho}_N, \boldsymbol{\eta}) \mathbf{a} = N^{1-2\alpha} \mathbf{a}^\top A(\mathbf{h}) \mathbf{a} + \mathbf{a}^\top B \mathbf{a} + o(N^{1-2\alpha}) + o(1). \quad (42)$$

Hence, in directions for which $\mathbf{a}^\top A(\mathbf{h}) \mathbf{a} > 0$ and $\mathbf{a}^\top B \mathbf{a} > 0$,

$$\|\boldsymbol{\rho}_N - \boldsymbol{\pi}\| \asymp N^{-1/2} \quad (43)$$

is the scale at which the mean and fluctuation channels contribute at the same order.

Proof. See Appendix H. □

Thus $N^{-1/2}$ is the crossover scale: above it the mean channel can dominate, at it the two nonzero channels are of the same asymptotic order, and below it the mean contribution vanishes while any nonzero fluctuation contribution remains $O(1)$.

At exact aggregate equilibrium, $\mathbf{m}_r = \mathbf{0}$ for every N , so within the Gaussian approximation the per-transition parameter information is

$$\mathcal{I}_{rs}^{(N)}(\boldsymbol{\pi}, \boldsymbol{\eta}) = B_{rs}, \quad (44)$$

independent of N . Thus a larger equilibrium population makes the observed proportions more concentrated but does not make a single aggregate transition arbitrarily informative about equilibrium-preserving changes in the transition dynamics. Information must instead accumulate through repeated fluctuation realizations across time or replicated populations. Under reversible continuous-time dynamics, Theorem 2 shows that this second-order channel is sufficiently rich to identify the full transition mechanism at the population level.

There is also a direct connection between this local scale and ordinary stationary behavior. If the N independent units are marginally stationary with distribution $\boldsymbol{\pi}$, then at any fixed observation time

$$\mathbf{Y}_t \sim \text{Multinomial}(N, \boldsymbol{\pi}),$$

and therefore

$$\frac{\mathbf{Y}_t}{N} - \boldsymbol{\pi} = O_p(N^{-1/2}). \quad (45)$$

Thus the critical regime in Theorem 3 is not merely an artificial local sequence: finite-population stationary fluctuations naturally occur on the same $N^{-1/2}$ scale at which mean and fluctuation information coexist.

Together, Theorems 2 and 3 describe how transition information persists at and near equilibrium. The first gives a constructive recovery result from equilibrium conditional covariance under reversible continuous-time dynamics. The second, without requiring reversibility, quantifies how the relative contribution of conditional-mean and conditional-covariance sensitivity changes as the observed source composition approaches equilibrium.

5 Numerical Studies

For the main numerical studies, we specialize the general CTMC framework to a reversible nearest-neighbor model. Let $\mathcal{E}_{\text{nn}} = \{\{j, j+1\} : j = 1, \dots, K-1\}$ be the path graph. For $\boldsymbol{\pi} \in \text{int}(\Delta_{K-1})$ and positive mobility parameters $\boldsymbol{\mu} = (\mu_1, \dots, \mu_{K-1})$, define

$$q_{j,j+1} = \mu_j \sqrt{\pi_{j+1}/\pi_j}, \quad q_{j+1,j} = \mu_j \sqrt{\pi_j/\pi_{j+1}},$$

with all other off-diagonal entries zero and diagonal entries chosen so rows sum to zero. This construction is reversible with stationary distribution $\boldsymbol{\pi}$. The parameterization follows from the standard reversible edge-flow representation and separates the equilibrium weighting from a symmetric coupling strength; its derivation and interpretation are given in Appendix I. For observation interval Δ , $\mathbf{P}(\boldsymbol{\pi}, \boldsymbol{\mu}) = \exp\{\Delta \mathbf{Q}(\boldsymbol{\pi}, \boldsymbol{\mu})\}$. Although $\mathbf{Q}$ is nearest-neighbor, $\mathbf{P}$ generally has positive non-neighboring entries because multiple continuous-time jumps can occur within one observation interval. In the likelihood experiments below, the general likelihoods of Section 3 are evaluated on this restricted parametric family, so $\hat{\mathbf{Q}} = \mathbf{Q}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}})$ and $\hat{\mathbf{P}} = \exp\{\Delta \hat{\mathbf{Q}}\}$.

The numerical studies are organized around five complementary questions. Experiment 1 compares exact and Gaussian likelihood inference and examines how estimation quality depends on population size, observation length, temporal resolution, and heterogeneity across transition mechanisms. Experiment 2 asks whether the equilibrium recovery result is visible in finite samples and how its numerical stability depends on the speed of the dynamics. Experiment 3 examines the near-equilibrium regime and tracks the predicted transfer of information from conditional-mean redistribution to finite-population fluctuations. Experiment 4 studies scalability in substantially larger histogram state spaces and separates the effects of population size, state-space dimension, and observation length. Experiment 5 uses a matched path-versus-nonlocal-tree comparison to examine whether the main estimation behavior persists when direct transitions are not restricted to neighboring states. Together, the experiments distinguish likelihood approximation, structural recoverability, finite-sample resolution, computational scalability, and sensitivity to transition topology.

Unless otherwise stated, we use

$$K = 4, \quad \Delta = 1, \quad \boldsymbol{\pi} = (0.12, 0.28, 0.38, 0.22), \quad \boldsymbol{\mu} = (0.35, 0.25, 0.30). \quad (46)$$

All histograms are generated from the exact unit-level CTMC model and then stripped of individual identities and transition tables before inference. The baseline values imply the generator $\mathbf{Q} = \mathbf{Q}(\boldsymbol{\pi}, \boldsymbol{\mu})$ and one-step transition matrix $\mathbf{P} = \exp(\mathbf{Q})$ shown in Figure 1. The generator is nearest-neighbor, whereas $\mathbf{P}$ has positive nonadjacent transition probabilities because multiple continuous-time jumps may occur within one observation interval.

5.1 Exact likelihood versus Gaussian approximation

Experiment 1 asks when the Gaussian likelihood is an adequate substitute for the exact finite-population likelihood, and which remaining errors reflect intrinsic resolution rather than approximation.

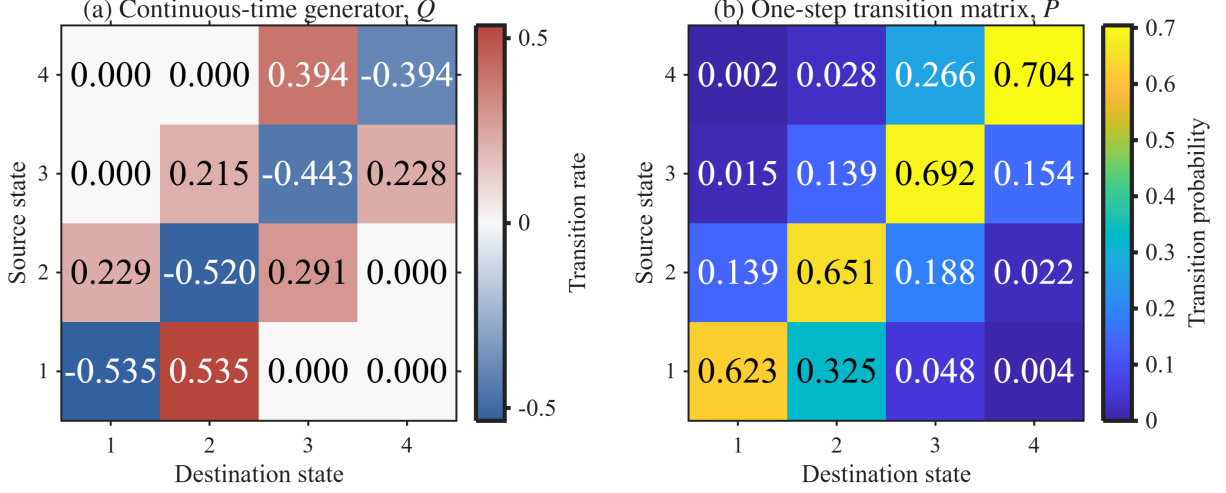


Figure 1: Baseline transition mechanism used in the numerical studies. Panel (a) shows the reversible nearest-neighbor generator $\mathbf{Q}(\boldsymbol{\pi}, \boldsymbol{\mu})$ for (46); panel (b) shows the corresponding one-step transition matrix $\mathbf{P} = \exp(\mathbf{Q})$ for $\Delta = 1$. Nonadjacent entries of $\mathbf{P}$ are positive because more than one continuous-time jump can occur between observations.

5.1.1 Population-size sensitivity

For this population-size comparison, the reference mechanism differs from (46):

$$\boldsymbol{\pi} = (0.26660926, 0.13909125, 0.35642353, 0.23787596), \quad \boldsymbol{\mu} = (0.25738153, 0.46589071, 0.18508825).$$

This is the fixed reference mechanism used for Figure 2. We first vary N at fixed T and Δ . This simultaneously tests the quality of the large-population Gaussian approximation and whether increasing population size improves recovery of the underlying dynamics. We compare the exact observed-data likelihood in (7) with the Gaussian approximation in (19). Two target source-composition regimes are considered: an equilibrium target, $\boldsymbol{\rho} = \boldsymbol{\pi}$, and a moderate nonequilibrium target,

$$\boldsymbol{\rho} = \boldsymbol{\pi} + 0.5\mathbf{h}, \quad \mathbf{h} = (0.12, -0.06, -0.04, -0.02). \quad (47)$$

The initial histogram is constructed by the largest-remainder rule: set $y_{0,j} = \lfloor N\rho_j \rfloor$ and allocate the remaining $N - \sum_j \lfloor N\rho_j \rfloor$ units to the largest fractional parts, resolving ties by state order. Thus $\|\mathbf{Y}_0/N - \boldsymbol{\rho}\|_\infty \leq N^{-1}$. The equilibrium-target initialization is generally only approximately at aggregate equilibrium; it is neither an exact fixed histogram $N\boldsymbol{\pi}$ nor a random multinomial stationary initialization. At the smallest N , rounding can materially change the intended departure from equilibrium or make two target regimes coincide. Subsequent histograms evolve without resetting.

For each regime we use

$$N \in \{3, 5, 7, 10, 20, 30, 40\}, \quad T = 20,$$

with $R = 200$ independent replications per population size and regime, giving 2800 fitted data sets. The values $N = 3, 5, 7$ deliberately stress the Gaussian approximation in a sparse-count setting. The exact likelihood is evaluated using the cached dynamic program in Appendix D; both methods use the same unconstrained parameterization and matched multistart optimization. Throughout the numerical comparisons, estimates denote the returned likelihood-based fits. Agreement between these fits does not certify attainment of either unrestricted global maximum.

For the fixed reference mechanism, estimation accuracy is summarized by

$$E_\pi = \|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2 / \sqrt{K}, \quad E_P = \|\hat{\mathbf{P}} - \mathbf{P}\|_F / K, \quad E_\mu = \frac{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2}{\|\boldsymbol{\mu}\|_2}. \quad (48)$$

For this population-size plot, the reported π and $\mathbf{P}$ errors are the root mean squared coordinate and entry errors, respectively: the stored squared norms are divided by K and K^2 before taking square roots. Mobility error is relative to the norm of the true mobility vector. The observation-length comparison below instead reports unscaled Euclidean and Frobenius errors for π and $\mathbf{P}$; numerical magnitudes across these panels should therefore not be compared directly. Medians are taken over the pooled equilibrium-target and moderate-nonequilibrium-target replications at each N .

Figure 2 shows a clear separation between *approximation error* and *intrinsic estimation difficulty*. At $N = 3$, the Gaussian approximation has a visible penalty: the pooled median error for π is 0.1735 compared with 0.1235 under the exact likelihood, and the corresponding errors for $\mathbf{P}$ are 0.1184 and 0.0952. The gap closes rapidly. By $N = 40$, the median errors are 0.0340 versus 0.0334 for π and 0.0520 versus 0.0508 for $\mathbf{P}$. Thus, once the population is no longer extremely small, the discrepancy between Gaussian and exact inference is minor relative to the underlying estimation error.

The mobility parameter reveals a second phenomenon. The pooled median relative error in μ falls from 0.583 under exact inference and 0.824 under Gaussian inference at $N = 3$ to 0.442 and 0.451, respectively, at $N = 20$. Beyond this point both curves flatten, reaching 0.424 and 0.427 at $N = 40$. Increasing N therefore removes most of the Gaussian-approximation penalty, but the mobility error levels off at fixed T . This indicates that the remaining difficulty is not primarily caused by the Gaussian approximation. Figure 2(d) makes the associated tail behavior explicit. At $N = 3$, 24.25% of Gaussian fits satisfy $\max_j \hat{\mu}_j > 10$, compared with 8.00% under the exact likelihood. This excess Gaussian instability decreases quickly with population size: at $N = 20$ the corresponding proportions are 4.75% and 4.25%, and at $N = 40$ they are 4.25% and 6.25%. Thus large mobility estimates persist under both likelihoods once the sparse-count approximation penalty has largely disappeared.

Coordinate-wise likelihood profiles clarify the origin of these estimates. In a representative fast-interface case, the exact relative log-likelihood changes by only 8.4×10^{-7} over $10^6 \leq \mu_1 \leq 10^{12}$, while the Gaussian profile changes by only 1.5×10^{-5} . The large estimates are therefore consistent with effectively flat fast-dynamics likelihood tails, rather than failed structural identification or a defect specific to the Gaussian approximation. Median rather than root-mean-square mobility error is used for this reason. These coordinate-wise evaluations do not establish that a finite global maximizer exists; a nearly flat tail can also make the reported numerical fit sensitive to stopping tolerances or parameter limits.

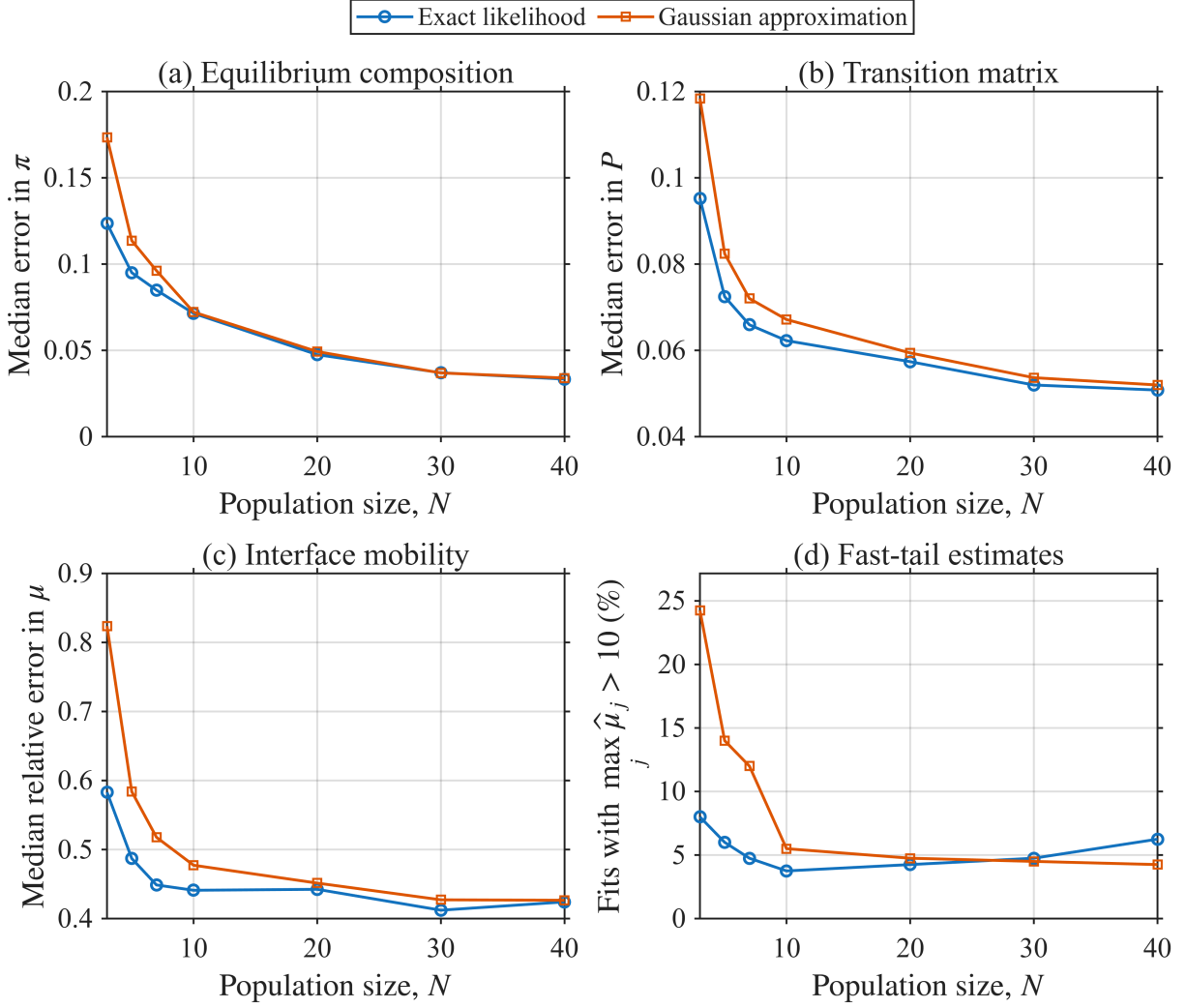


Figure 2: Experiment 1: exact finite-population likelihood versus the Gaussian approximation. Panels (a)-(c) show pooled median estimation errors, with the coordinate and entry normalizations in (48), for the equilibrium composition π , one-step transition matrix $\mathbf{P}$, and mobility vector μ over the two rounded target regimes. Panel (d) shows the percentage of fitted data sets with $\max_j \hat{\mu}_j > 10$. The Gaussian approximation incurs a visible penalty at very small N , but its estimates rapidly approach those of the exact likelihood as N increases. Large mobility estimates occur under both likelihoods, consistent with nearly flat likelihood tails toward very large candidate mobility values at fixed T and Δ .

5.1.2 Observation-length sensitivity

The preceding results suggest that increasing N primarily improves the Gaussian approximation, while residual uncertainty in the continuous-time mobility parameters remains at fixed observation length. We therefore hold $N = 20$ and $\Delta = 1$ and vary T to determine whether additional temporal observations resolve this uncertainty:

$$T \in \{10, 20, 50, 100, 200\}.$$

Here the mechanism is the baseline in (46). We consider the two target initial compositions $\rho_0 = \pi$ and $\rho_0 = \pi + 0.5\mathbf{h}$, converted to integer histograms by the largest-remainder rule above, with $R = 100$ replications per initialization regime. Within each replication, a trajectory of length $T = 200$ is generated and the shorter data sets are obtained from prefixes of the same

trajectory, so comparisons across observation lengths are paired. Exact and Gaussian likelihood fits again use matched multistart optimization.

Figure 3 shows a systematic improvement in estimation accuracy as the observation length increases. Under equilibrium-target initialization, the median exact-likelihood errors in $\boldsymbol{\pi}$, $\mathbf{P}$, and $\boldsymbol{\mu}$ decrease from 0.123, 0.356, and 0.561, respectively, at $T = 10$, to 0.022, 0.065, and 0.114 at $T = 200$. The moderate-nonequilibrium regime exhibits a similar decline. Thus, although increasing N at fixed T leaves a residual mobility-resolution limit in Figure 2, increasing the number of observed transitions substantially improves recovery of both the stationary composition and the transition dynamics.

The Gaussian approximation closely tracks the exact likelihood throughout this experiment. The discrepancy is most visible for the shortest sequences and becomes small relative to the estimation error as T increases. For example, under equilibrium-target initialization at $T = 200$, the median errors in $\mathbf{P}$ are 0.065 and 0.068 for the exact and Gaussian likelihoods, respectively, while the corresponding relative errors in $\boldsymbol{\mu}$ are 0.114 and 0.122.

Longer observation sequences also remove most of the extreme mobility estimates seen for short trajectories. Under equilibrium-target initialization, the fraction of exact-likelihood fits satisfying $\max_j \hat{\mu}_j > 10$ falls from 14% at $T = 10$ to zero by $T = 100$, with essentially the same pattern for the Gaussian likelihood. This supports the interpretation that the large mobility estimates observed for short trajectories primarily reflect limited information from the number of observed transitions rather than a systematic failure of either likelihood approximation.

Together, the population-size and observation-length experiments distinguish two effects: increasing N primarily improves the Gaussian approximation to the finite-population likelihood, whereas increasing T provides additional observed transitions and continues to improve recovery of the transition dynamics.

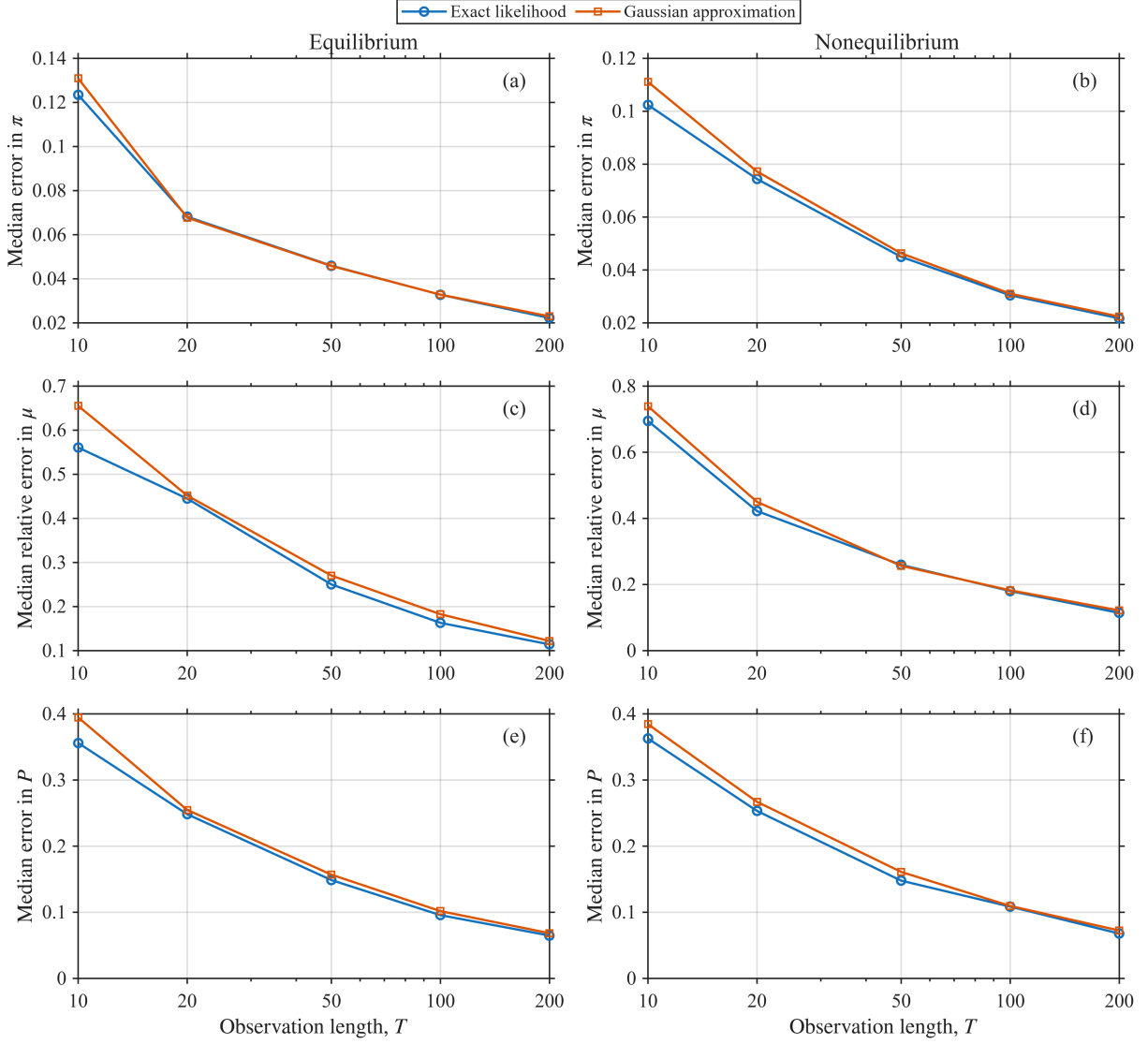


Figure 3: Observation-length sensitivity in Experiment 1. The population size and observation interval are fixed at $N = 20$ and $\Delta = 1$, while the observation length T is varied. The left and right columns correspond to equilibrium-target and moderate-nonequilibrium-target initialization, respectively. The rows show median estimation errors for the stationary composition π , mobility vector μ , and one-step transition matrix P . Estimation accuracy improves systematically with observation length under both source regimes, while the Gaussian approximation closely tracks the exact finite-population likelihood.

5.1.3 Observation-interval sensitivity

Population size and observation length are not the only determinants of recoverability. Even with sufficient observations, continuous-time rates can be difficult to recover when the sampling interval is long relative to the dynamics. We therefore hold $N = 20$ and $T = 20$ fixed and vary only the observation interval,

$$\Delta \in \{0.25, 0.5, 1, 2\}.$$

For each value of Δ , we generate $R = 100$ independent trajectories with the rounded equilibrium-target initialization under the same baseline generator in (46) and fit both likelihoods using matched multistart optimization. Because the number of observed transitions T is held fixed, the total elapsed time $T\Delta$ varies with Δ ; the diagnostic therefore examines the effect of spacing

between successive observations rather than a fixed-horizon sampling design. Figure 4 shows the resulting mobility error and the frequency of large mobility estimates. The effect of the observation interval is pronounced. Under the exact likelihood, the median relative mobility error increases from 0.358 at $\Delta = 0.25$ to 0.798 at $\Delta = 2$; under the Gaussian approximation it increases from 0.374 to 0.947. More strikingly, the proportion of fits satisfying $\max_j \hat{\mu}_j > 10$ rises from 0% at $\Delta = 0.25$ to 21% at $\Delta = 2$ for the exact likelihood and from 0% to 24% for the Gaussian approximation. At the intermediate values $\Delta = 0.5$ and 1, the corresponding exact/Gaussian frequencies are 1%/1% and 5%/4%. The two methods therefore exhibit essentially the same rise in the frequency of large mobility estimates as the observation interval becomes coarser.

The same pattern is visible on the finite-time transition scale: the median relative error in $\mathbf{P}$ increases from 0.061 to 0.244 under the exact likelihood and from 0.062 to 0.243 under the Gaussian approximation as Δ increases from 0.25 to 2. Because $\mathbf{P} = \exp\{\Delta \mathbf{Q}\}$ changes with Δ , the transition-matrix errors at different observation intervals correspond to different finite-time estimands. The mobility vector, by contrast, is fixed across the experiment. Together with the population-size and observation-length experiments, the results show that the large mobility estimates are not primarily an artifact of the Gaussian approximation: they become less frequent with additional observed transitions and more frequent as the observation interval increases, consistent with weak finite-sample resolution of continuous-time rates when observations provide insufficient information about the underlying time scale.

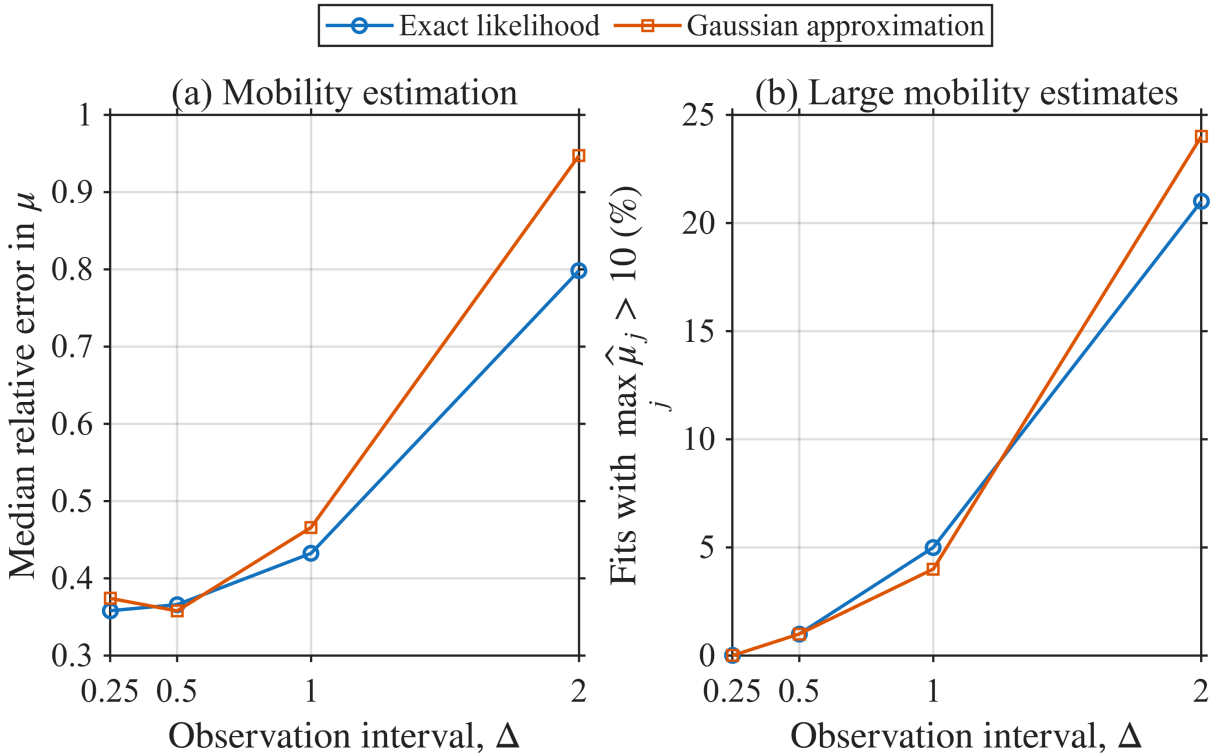


Figure 4: Observation-interval sensitivity in Experiment 1. The population size and observation length are fixed at $N = 20$ and $T = 20$, while the observation interval Δ is varied. Panel (a) shows the median relative mobility error and panel (b) shows the percentage of fits satisfying $\max_j \hat{\mu}_j > 10$. Both exact and Gaussian inference deteriorate as observations become more widely spaced, and the frequency of large mobility estimates rises in nearly the same way under the two likelihoods. The similar behavior of the two likelihoods indicates that the deterioration is not specific to the Gaussian approximation.

Because performance under one transition mechanism need not reflect the range of finite-sample difficulties encountered across the parameter space, we next repeat the comparison over

a heterogeneous collection of mechanisms.

5.1.4 Performance across heterogeneous transition mechanisms

The preceding comparisons use a single baseline transition mechanism. To check that the observed agreement between exact and Gaussian inference is not peculiar to that mechanism, we repeat the comparison over 100 distinct transition mechanisms. We retain $K = 4$, $\Delta = 1$, and $T = 20$, but generate a new interior stationary distribution and mobility vector for each mechanism. Stationary distributions are drawn from a symmetric Dirichlet distribution with concentration parameter 4, subject to $0.05 \leq \pi_j \leq 0.70$, and the mobility coordinates are drawn independently on a log scale over $[0.12, 0.60]$. Half of the mechanisms use an equilibrium target and half use a moderate nonequilibrium target; both are converted to integer source histograms by the same rounding rule. The same bank of 100 mechanisms is used at $N \in \{5, 20, 40\}$, and exact and Gaussian fits use matched optimization starts. This experiment deliberately exposes the two methods to mechanisms of heterogeneous difficulty rather than repeatedly sampling from a single transition system.

Figure 5 separates two questions. The left column measures recovery of the true finite-time transition matrix and mobility vector, whereas the right column measures the additional discrepancy incurred by replacing the exact likelihood with its Gaussian approximation. Points are medians across the 100 mechanisms and vertical bars are interquartile ranges; the bars therefore describe heterogeneity in mechanism difficulty, not Monte Carlo uncertainty in the medians.

The heterogeneous study is substantially more challenging than the reference configuration. At $N = 40$, the median relative mobility error is 0.464 under the exact likelihood and 0.469 under the Gaussian approximation. Nevertheless, the median relative Gaussian-exact discrepancy in mobility is only 0.059. For the one-step transition matrix, the corresponding errors relative to truth are 0.168 and 0.183, while the median Gaussian-exact discrepancy is only 0.018. Across $N = 5, 20, 40$, the latter discrepancy decreases from 0.317 to 0.107 to 0.059 for $\boldsymbol{\mu}$, and from 0.090 to 0.036 to 0.018 for $\mathbf{P}$. Thus mechanisms can be difficult to estimate even under the exact finite-population likelihood while the Gaussian fit remains close to the exact fit.

The small increase in the marginal median errors from $N = 20$ to $N = 40$ does not indicate systematic deterioration with population size. In paired mechanism-by-mechanism comparisons, the exact mobility error improves for 50% of mechanisms and worsens for 50%, with a median paired change of only 0.0019; the corresponding median change for $\mathbf{P}$ is 0.0031. Increasing N from 5 to 20, by contrast, produces a clear improvement. Thus, across this heterogeneous collection of mechanisms, increasing population size beyond moderate levels provides little additional improvement at fixed T and Δ , consistent with the plateau observed for the baseline mechanism. Large mobility estimates also occur under both likelihoods: the proportions of fits with $\max_j \hat{\mu}_j > 10$ are 9%, 9%, and 7% for the exact likelihood and 16%, 9%, and 7% for the Gaussian approximation at $N = 5, 20, 40$, respectively. Their common occurrence at moderate and large N is consistent with the flat fast-dynamics likelihood profiles observed in the reference study rather than an approximation-specific failure.

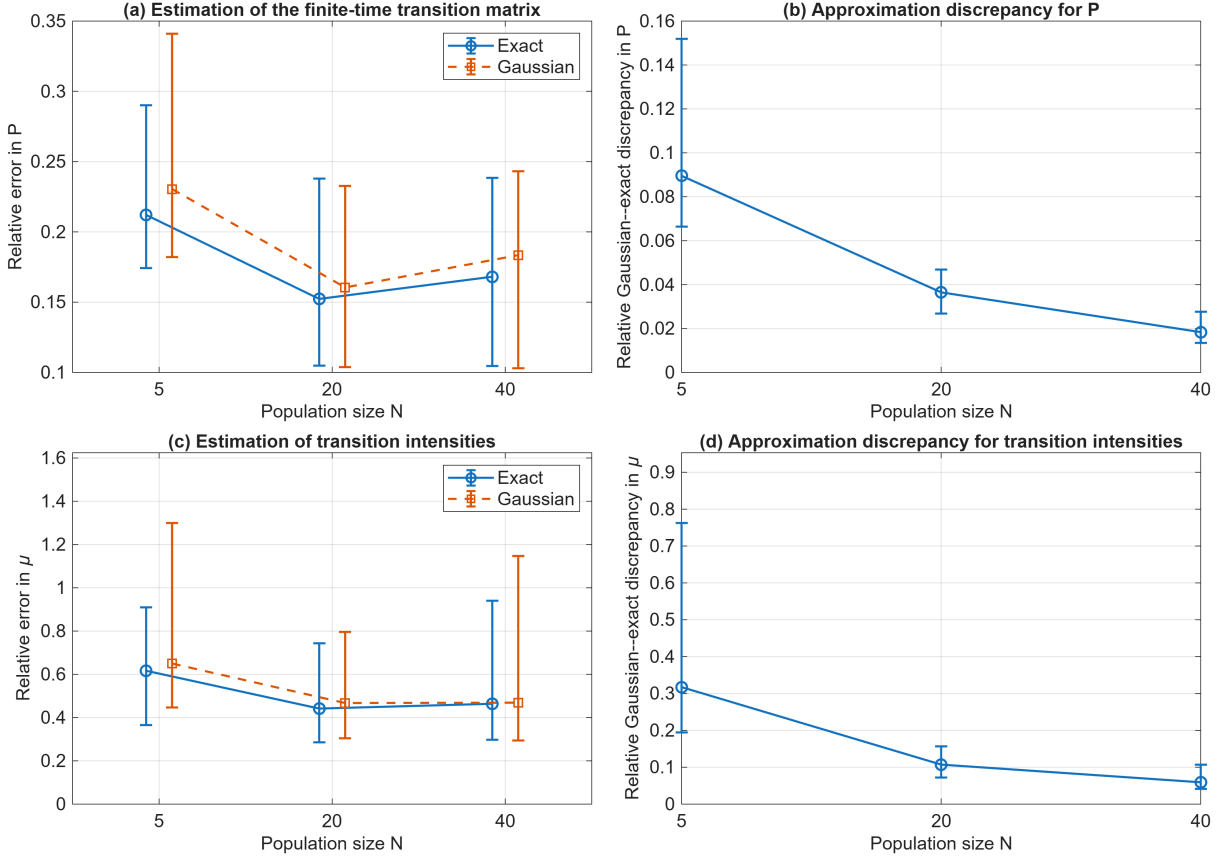


Figure 5: Experiment 1 robustness across 100 heterogeneous transition mechanisms. Panels (a) and (c) show relative errors with respect to the true one-step transition matrix $\mathbf{P}$ and mobility vector μ for the exact and Gaussian likelihoods. Panels (b) and (d) show the corresponding relative Gaussian-exact discrepancies. Points denote medians across the 100 mechanisms and vertical bars denote interquartile ranges, thereby displaying heterogeneity across transition mechanisms rather than sampling uncertainty in the median. Although recovery of the true dynamics can remain difficult at fixed $T = 20$, particularly for μ , the discrepancy between Gaussian and exact inference decreases markedly with population size.

5.2 Recovery from equilibrium fluctuations

Experiment 2 asks whether the constructive equilibrium inversion is visible at finite sample sizes, and how its stability changes as the dynamics become faster. We fix $N = 1000$ and consider three mobility levels,

$$\mu^{(c)} = c\mu, \quad c \in \{0.6, 1, 1.5\}, \quad (49)$$

representing slower, baseline, and faster dynamics. For each mobility level and each of $R = 100$ Monte Carlo replications, we generate 5000 independent one-step transitions conditional on the same exact equilibrium source histogram $\mathbf{Y}_{t-1} = N\pi$. Let $\mathbf{Y}^{(\ell)}$ denote the destination histogram in the ℓ th such transition, and let L denote the number of replicated equilibrium transitions used to estimate the conditional covariance. We consider nested sample sizes

$$L \in \{50, 100, 250, 500, 1000, 2000, 5000\}$$

using prefixes of the same 5000 transitions within each Monte Carlo replication. Because the experiment conditions on the known equilibrium composition π , covariance is computed about

the population conditional mean $N\boldsymbol{\pi}$:

$$\hat{\boldsymbol{\Sigma}}_{\pi,L} = \frac{1}{L} \sum_{\ell=1}^L (\mathbf{Y}^{(\ell)} - N\boldsymbol{\pi})^\top (\mathbf{Y}^{(\ell)} - N\boldsymbol{\pi}), \quad (50)$$

after which we apply a regularized spectral version of the population inversion. Symmetrize $\mathbf{I}_K - \hat{\mathbf{C}}_{\pi,L}$ and write its eigendecomposition as $\mathbf{U} \text{diag}(\hat{\lambda}_i) \mathbf{U}^\top$. The implementation uses

$$\tilde{\lambda}_i = \min\{1, \max\{10^{-10}, \hat{\lambda}_i\}\}, \quad \hat{\mathbf{H}}_L = \frac{1}{2\Delta} \mathbf{U} \text{diag}(\log \tilde{\lambda}_i) \mathbf{U}^\top,$$

$$\hat{\mathbf{P}}_L = \mathbf{D}_\pi^{-1/2} \mathbf{U} \text{diag}(\sqrt{\tilde{\lambda}_i}) \mathbf{U}^\top \mathbf{D}_\pi^{1/2}, \quad \hat{\mathbf{Q}}_L = \mathbf{D}_\pi^{-1/2} \hat{\mathbf{H}}_L \mathbf{D}_\pi^{1/2}, \quad \hat{\boldsymbol{\mu}}_{L,j} = (\hat{\mathbf{H}}_L)_{j,j+1}.$$

The eigenvalue floor makes these matrix functions finite, but does not impose nonnegative off-diagonal rates or the nearest-neighbor sparsity constraint. The reported errors include these unconstrained spectral reconstructions, without clipping negative extracted mobilities or discarding matrices that fail the generator constraints. Accordingly, $\hat{\mathbf{P}}_L$ and $\hat{\mathbf{Q}}_L$ need not be valid transition matrices and generators in finite samples. The experiment measures the numerical accuracy of a regularized inverse map; it is not an evaluation of a constrained CTMC estimator. The logarithm is sensitive to the chosen floor when clipping occurs, and the reported results do not establish robustness to that choice.

Figure 6 reports the median relative errors

$$E_P(L) = \frac{\|\hat{\mathbf{P}}_L - \mathbf{P}\|_F}{\|\mathbf{P}\|_F}, \quad E_\mu(L) = \frac{\|\hat{\boldsymbol{\mu}}_L - \boldsymbol{\mu}\|_2}{\|\boldsymbol{\mu}\|_2}. \quad (51)$$

This is deliberately a replicated conditional experiment: every transition starts from the same exact aggregate-equilibrium histogram $N\boldsymbol{\pi}$, and L controls the number of independent fluctuation realizations available for estimating $\boldsymbol{\Sigma}_\pi$. The experiment therefore isolates the inversion in Theorem 2; it is not intended as a covariance estimator for a single stationary trajectory. Recovery improves steadily with L . For $c = 0.6$, the median relative errors in $\mathbf{P}$ and $\boldsymbol{\mu}$ fall from 12.5% and 37.9% at $L = 50$ to 1.21% and 3.89% at $L = 5000$. For $c = 1$, they fall from 25.2% and 62.5% to 2.58% and 6.28%, respectively. Thus repeated equilibrium fluctuations support increasingly accurate recovery even though the conditional mean remains fixed.

The fast mechanism, $c = 1.5$, separates *recoverability* from *conditioning*. The median relative error in $\mathbf{P}$ improves from 31.4% at $L = 50$ to 4.31% at $L = 5000$, but the continuous-time mobility is much less stable for small L : its median relative error exceeds 400% for $L \leq 250$, falls to 40.2% at $L = 500$, and reaches 10.6% at $L = 5000$. Recovering $\boldsymbol{\mu}$ requires the matrix logarithm of an estimated second-order quantity, so errors that are modest on the finite-time transition scale can be amplified substantially when mapped back to continuous time.

Panel (c) shows that this is a systematic conditioning effect. Over $0.2 \leq c \leq 2$, $\lambda_{\min}(\mathbf{S}^2)$ decreases as mobility increases; at the three simulated scales,

$$\lambda_{\min}(\mathbf{S}^2) = 0.3122, \quad 0.1437, \quad 0.05445.$$

As $\mathbf{S}^2$ approaches singularity, the square-root and especially the logarithmic inverse become increasingly sensitive to covariance-estimation error. The three simulated mobility levels therefore represent points along a continuous deterioration in conditioning rather than isolated examples.

At the population level, however, the inversion remains exact. Substituting the exact equilibrium covariance into the recovery formulas reconstructs $\mathbf{P}$, $\mathbf{Q}$, and $\boldsymbol{\mu}$ to machine precision at all three mobility levels. Experiment 2 therefore illustrates the substantive content of Theorem 2: the population conditional covariance of equilibrium fluctuations, together with the observed equilibrium composition, retains enough information to recover the transition dynamics exactly, while finite-sample inversion becomes progressively harder as the dynamics become fast relative to the observation interval.

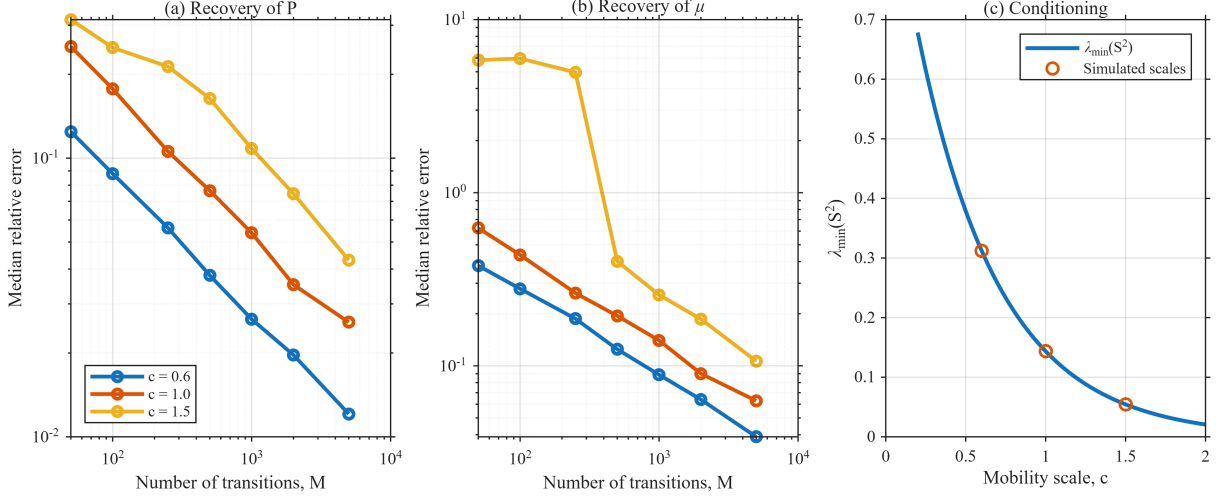


Figure 6: Experiment 2: recovery from repeated equilibrium fluctuations. Panels (a) and (b) show median relative errors in the recovered one-step transition matrix $\mathbf{P}$ and mobility vector $\boldsymbol{\mu}$ as the number L of independent equilibrium transitions increases for mobility scales $c = 0.6, 1$, and 1.5 . Panel (c) shows $\lambda_{\min}(\mathbf{S}^2)$ over a dense mobility-scale range, with the three simulated cases marked. Recovery improves with L , but the inverse problem becomes increasingly ill-conditioned as the dynamics become faster; this affects continuous-time mobility much more strongly than the finite-time transition matrix. Using the exact equilibrium covariance recovers all quantities to machine precision.

5.3 Near-equilibrium information transition

Experiment 3 asks whether the predicted transfer of information from the conditional mean to fluctuations is visible along finite- N local sequences. We evaluate the Gaussian Fisher-information decomposition from Theorem 3 along the local sequences

$$\boldsymbol{\rho}_N = \boldsymbol{\pi} + N^{-\alpha} \mathbf{h}, \quad \alpha \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}, \quad (52)$$

together with exact equilibrium, using the same departure direction $\mathbf{h}$ as in (47). We use

$$N \in \{50, 100, 200, 500, 1000, 2000, 5000, 10000, 20000\}.$$

The calculation is deterministic and uses the continuous extension of the Gaussian moment formulas to all simplex compositions; the displayed $N\boldsymbol{\rho}_N$ need not be integer-valued. For $0 < \alpha < 1$, as in this experiment, largest-remainder rounding changes $\boldsymbol{\rho}_N$ by $O(N^{-1}) = o(N^{-\alpha})$, so admissible rounded sequences have the same leading asymptotic expansion, although the plotted finite- N values use the unrounded compositions. The Gaussian Fisher-information decomposition is evaluated directly from the model moments along the scalar log-mobility path

$$\boldsymbol{\mu}(\beta) = e^\beta \boldsymbol{\mu}, \quad (53)$$

so that β represents a proportional change in the overall mobility level.

Along the scalar path (53), let $I_{\text{mean}}^{(N)}$ and $I_{\text{fluct}}^{(N)}$ denote, respectively, the directional contributions of the mean and fluctuation terms in (36). Figure 7 displays the fraction

$$\eta_N = \frac{I_{\text{mean}}^{(N)}}{I_{\text{mean}}^{(N)} + I_{\text{fluct}}^{(N)}}, \quad (54)$$

which directly measures where the information resides. Values near one indicate dominance of systematic mean redistribution; values near zero indicate dominance of fluctuations. The

predicted transition at $\alpha = 1/2$ is clear. For $\alpha = 0.2 < 1/2$, the mean-information share rises from 0.280 at $N = 50$ to 0.940 at $N = 20000$; at $\alpha = 0.5$ it remains essentially constant; and for $\alpha = 0.8 > 1/2$ it decreases toward zero. At exact equilibrium, the mean-information share is identically zero.

The asymptotic rates agree closely with Theorem 3. Using the largest five population sizes, the fitted log – log slopes of the mean-information term for $\alpha = 0.2, 0.3, \dots, 0.8$ are

$$0.611, 0.407, 0.204, 0.002, -0.199, -0.399, -0.600,$$

compared with the theoretical values

$$1 - 2\alpha = 0.6, 0.4, 0.2, 0, -0.2, -0.4, -0.6.$$

The fitted fluctuation-information slope is numerically zero throughout the grid. At exact equilibrium, the mean contribution vanishes while the fluctuation contribution remains approximately 0.464 for every N considered.

This experiment illustrates Theorem 3 along the single common log-mobility direction used here; it is not intended as a numerical validation of every parameter direction covered by the general theorem. The finite- N curves also show why the $N^{-1/2}$ boundary should be interpreted as an asymptotic information boundary rather than an abrupt finite-population switch. For example, $\alpha = 0.4 < 1/2$ is asymptotically mean-dominated, yet even at $N = 20000$ the mean contributes only about 0.235 of the total information. Close to the critical scaling, the eventual dominance of one channel can therefore emerge slowly, and fluctuation information may remain practically important even for large populations.

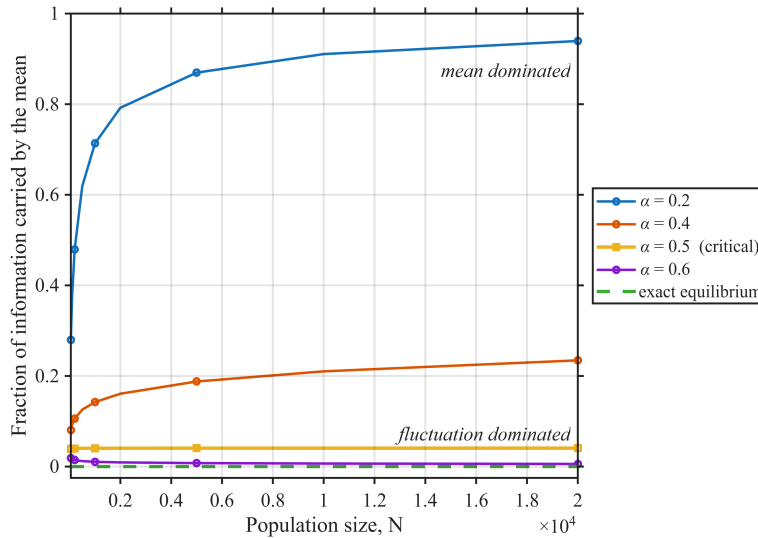


Figure 7: Experiment 3: transition in the information channel near equilibrium. The vertical axis is the fraction of Gaussian Fisher information about the common log-mobility direction carried by the conditional mean, $I_{\text{mean}}/(I_{\text{mean}} + I_{\text{fluct}})$. For $\alpha < 1/2$, this share increases with N ; at $\alpha = 1/2$, it remains approximately constant; and for $\alpha > 1/2$, it decreases toward zero. Exact equilibrium has zero mean-information share for every N . The figure therefore illustrates both the predicted $N^{-1/2}$ boundary and the slow finite- N emergence of the asymptotic regimes when α lies close to $1/2$.

5.4 Scalability in large histogram state spaces

Experiment 4 asks whether the Gaussian likelihood remains computationally and statistically useful when the histogram dimension is much larger than in the baseline examples. We consider

$$K \in \{4, 10, 20, 50, 100\},$$

so that the largest model contains 99 mobility parameters in addition to the equilibrium-composition parameters. All data are generated from the exact unit-level CTMC model; only the aggregate histograms are supplied to the estimator. The Gaussian likelihood is optimized using its analytic gradient. The purpose of the experiment is not to propose a new optimization method, but to distinguish three potential sources of difficulty: population size, state-space dimension, and observation length.

The computational motivation for using the Gaussian likelihood is already visible in the $K = 4$ reference problem. There, the cached exact likelihood is approximately 4.1-4.4 times slower than Gaussian optimization at $N = 10$, increasing to approximately 119-121 times slower at $N = 40$. We therefore use the Gaussian likelihood in the larger-dimensional experiments below.

Population size: we first fix $T = 100$ and vary the average occupancy per state,

$$N/K \in \{10, 50, 200, 1000\},$$

using $R = 50$ replications per configuration. Increasing the population size has only a modest effect on mobility estimation once the states are already reasonably populated. For example, at $K = 100$ the median relative mobility error is 0.200, 0.191, 0.187, and 0.189 as N/K increases from 10 to 1000. Across the same configurations no fitted mobility component exceeds 10. Thus, over the range studied, making the population much larger does not substantially improve resolution of the continuous-time transition intensities at fixed observation length.

State-space dimension: increasing K has a much clearer computational effect than statistical one. At $N/K = 10$, median fitting time increases from 0.16 seconds for $K = 4$ to 0.71 seconds for $K = 20$, 6.40 seconds for $K = 50$, and 40.9 seconds for $K = 100$. In contrast, the median relative mobility errors over the same sequence are 0.181, 0.186, 0.200, and 0.200 for $K = 4, 20, 50, 100$, respectively. The computational cost therefore grows strongly with state-space dimension, while the statistical error remains comparatively stable in this experiment. The weak dependence of runtime on N/K is also consistent with the aggregate Gaussian likelihood operating on a K -dimensional histogram rather than on individual trajectories.

Observation length: we isolate the effect of observation length by fixing $N/K = 10$ and considering

$$K \in \{20, 50, 100\}, \quad T \in \{20, 50, 100, 200, 500\},$$

again with $R = 50$ replications per configuration. From $T = 50$ onward the three error curves are nearly indistinguishable. At $T = 500$, the median relative mobility errors are 0.080, 0.082, and 0.085 for $K = 20, 50, 100$, respectively, while the corresponding median relative errors in the one-step transition matrix are 0.028, 0.029, and 0.031. Over the larger values of T , the empirical decay of the median mobility error is close to a $T^{-1/2}$ reference rate; the fitted slopes for $K = 20$ and $K = 50$ are -0.552 and -0.543 , respectively. This reference is descriptive rather than a claim of a new asymptotic theorem.

Very short sequences reveal a different phenomenon. At $T = 20$, a small but non-negligible fraction of fits contains a very large mobility component: 4%, 16%, and 26% for $K = 20, 50, 100$, respectively. The corresponding transition-matrix errors remain much more moderate. By $T = 50$, no fitted mobility component exceeds 10 in any of the simulated configurations. This mirrors Experiment 1: with insufficient temporal information, the finite-time transition matrix can be estimated substantially more stably than the continuous-time rates that generate it.

Figure 8 summarizes the experiment. Panel (a) shows a representative high-dimensional reconstruction at $K = 100$: the median estimated mobility profile closely tracks the true profile across all 99 interfaces. Panel (b) shows the rapid improvement with observation length and

the near-collapse of the $K = 20, 50, 100$ curves once T is at least 50. Panel (c) shows that computation is driven primarily by the number of states rather than by the population size. Taken together, the experiments show that increasing population size mainly improves aggregate-count precision, whereas increasing observation length has a much stronger effect on resolution of the continuous-time dynamics once states are reasonably populated.

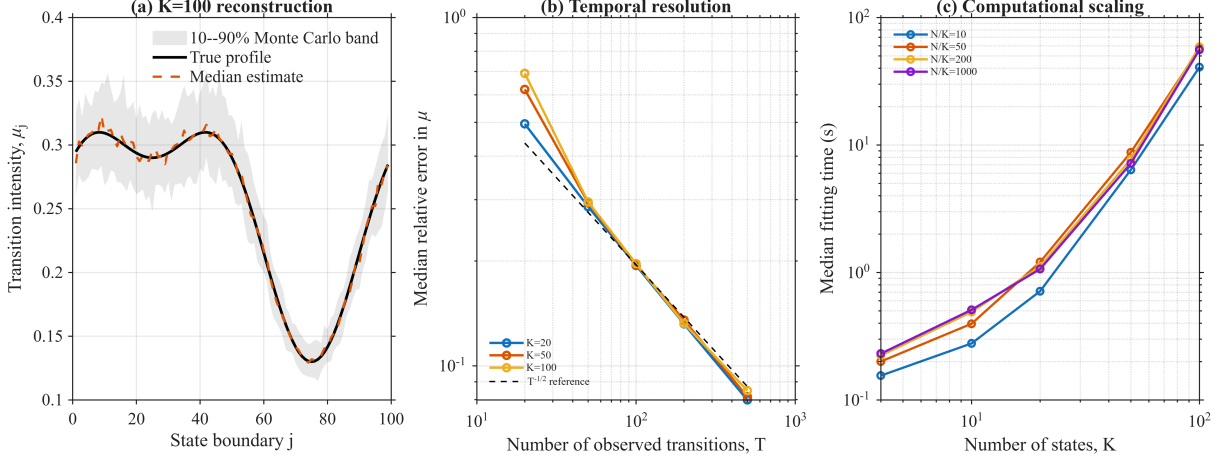


Figure 8: Experiment 4: scalability of Gaussian aggregate inference. Panel (a) shows reconstruction of the $K = 100$ mobility profile, including the true profile, pointwise median estimate, and 10-90% Monte Carlo band. Panel (b) shows median relative mobility error as a function of the number T of observed transitions for $K = 20, 50, 100$ with $N/K = 10$; from $T = 50$ onward the curves are nearly insensitive to K and decrease approximately in line with the displayed $T^{-1/2}$ reference rate. Panel (c) shows median fitting time as a function of K for several average occupancies N/K at $T = 100$. Computational cost increases strongly with state-space dimension but only weakly with population size.

5.5 Robustness to transition topology

The preceding experiments use a nearest-neighbor generator as a controlled and parsimonious specialization of the general model. To assess whether the main estimation behavior depends on this path structure, we repeat the observation-length comparison using two reversible models with $K = 10$ states and the same number of mobility parameters. The *nearest-neighbor* model uses the path graph above, with direct transitions only between states j and $j + 1$. The *nonlocal* model uses a connected tree with the same $K - 1 = 9$ edges, but direct transitions are not restricted to adjacent state labels. Thus the two models have the same parameter dimension and differ only in the pattern of allowed direct transitions. The nonlocal tree uses the edge set

$$\mathcal{E}_{\text{tree}} = \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{2, 5\}, \{3, 6\}, \{3, 7\}, \{4, 8\}, \{5, 9\}, \{7, 10\}\}.$$

It has branching vertices and is not a relabeling of the path.

For each undirected edge $\{j, k\}$, the nonlocal model uses the same reversible edge parameterization,

$$q_{jk} = \mu_{jk} \sqrt{\pi_k / \pi_j}, \quad q_{kj} = \mu_{jk} \sqrt{\pi_j / \pi_k},$$

with all other off-diagonal entries zero and diagonal entries chosen so that rows sum to zero. We use $N = 100$, $\Delta = 1$,

$$T \in \{20, 50, 100, 200\},$$

and $R = 50$ independent replications for each transition structure. Estimation uses the Gaussian likelihood.

Figure 9 shows the resulting errors. The two transition structures exhibit very similar behavior, particularly once $T \geq 50$. At $T = 200$, the median transition-matrix errors are 0.045 and 0.044 for the nearest-neighbor and nonlocal models, respectively, while the corresponding median relative mobility errors are 0.126 and 0.127. Increasing observation length therefore produces essentially the same improvement under the two structures. Within this controlled comparison, the principal numerical conclusions are not specific to nearest-neighbor movement.

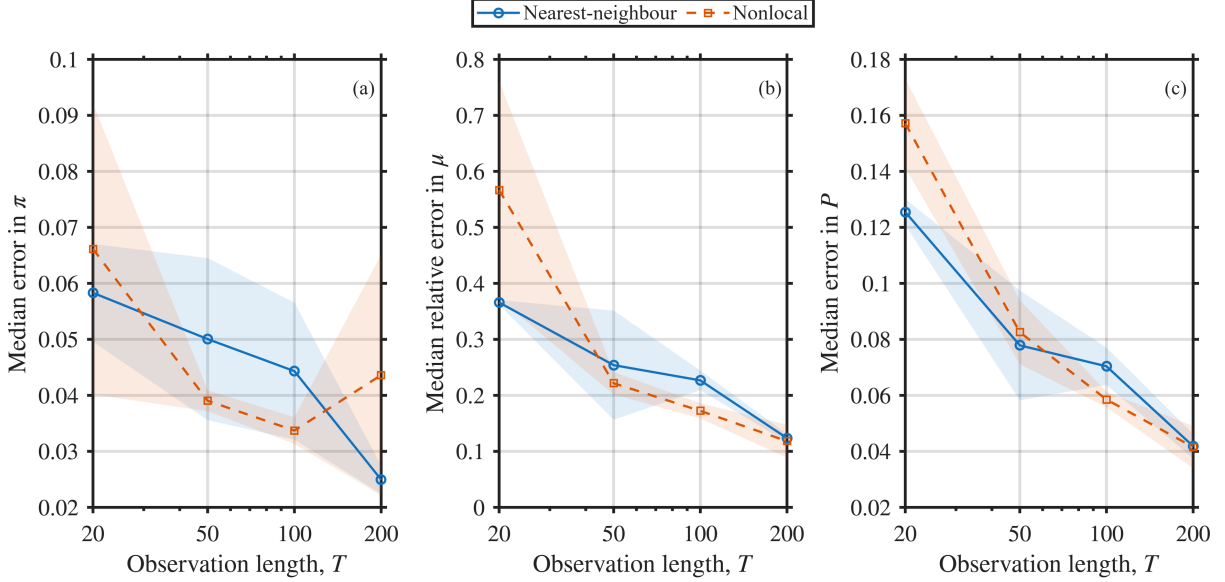


Figure 9: Robustness to transition topology. The nearest-neighbor model permits direct transitions only between adjacent states. The nonlocal model uses a connected tree with the same $K - 1$ edges, and hence the same number of mobility parameters, but allows direct transitions between non-adjacent states. Panels (a)-(c) show E_π , E_μ , and E_P , respectively, as functions of the number T of observed transitions. Curves show medians and shaded regions show the interquartile ranges.

5.6 Summary of results

The numerical studies distinguish likelihood approximation from statistical information about the transition dynamics. Increasing population size rapidly reduces the discrepancy between the Gaussian and exact likelihood estimators, but at fixed observation length this does not necessarily produce a comparable improvement in recovery of the underlying continuous-time mobility parameters. Increasing T , by contrast, supplies additional observed transitions and continues to improve recovery. The observation interval plays a different role again: through $\mathbf{P} = \exp\{\Delta\mathbf{Q}\}$ it determines how the continuous-time generator is encoded in each finite-time transition. Population size, observation length, and observation interval are therefore not interchangeable sources of information. Across the heterogeneous collection of mechanisms, difficult cases occur under both the exact and Gaussian likelihoods, indicating that finite-data conditioning, rather than Gaussian approximation alone, can limit recovery.

The equilibrium experiments illustrate the theoretical distinction between mean and fluctuation information. Repeated transitions from an exact equilibrium source permit increasingly accurate reconstruction from the conditional covariance, as predicted by Theorem 2, although structural recoverability does not imply that the finite-sample inversion is well conditioned. The near-equilibrium Fisher-information calculations illustrate the $N^{-1/2}$ transition of Theorem 3: the conditional-mean contribution grows for departures larger than this scale in the parameter direction studied, is of the same asymptotic order as fluctuation information at the critical scale,

and vanishes for smaller departures, while the fluctuation contribution remains $O(1)$.

Finally, the Gaussian approach remains computationally feasible over the state-space sizes examined here, up to $K = 100$. Its computational cost increases substantially with K but only weakly with population size once the required matrices have been formed. The matched path-versus-nonlocal-tree comparison shows similar recovery behavior for the two equally sparse reversible transition structures considered, particularly for moderate and large T ; it should not be interpreted as topology independence more generally.

6 Discussion

The principal result identifies a transition mechanism from the destination covariance conditional on an equilibrium source. Its statistical observable differs from stationary lagged cross-covariance: conditioning removes source variability, while reversibility and continuous time make the remaining covariance invertible through a matrix square root and logarithm. This establishes a route from a conditional second-order law to the generator, complementing stationary aggregate-moment recovery.

The three theoretical results answer different questions. Theorem 1 supplies sufficient conditions for recovering labeled transition probabilities from known conditional histogram laws. Theorem 2 gives a constructive moment inversion within the reversible CTMC class. Theorem 3 quantifies directional sensitivity within a Gaussian conditional model. None of these statements alone establishes consistency, asymptotic normality, or finite-sample accuracy of a numerical likelihood fit.

The distinction matters when interpreting the numerical evidence. The covariance-recovery experiment uses replicated transitions from a fixed equilibrium source, and the information experiment evaluates the Gaussian information formula deterministically. The likelihood comparisons instead fit histogram sequences with varying source compositions and jointly estimate stationary weights and mobilities. Their observed mobility-error plateaus are consistent with the information analysis, but are not a direct consequence of the fixed-composition, equilibrium-preserving theorem. Likewise, agreement between exact and Gaussian fits in the studied configurations is empirical evidence about the approximation; it does not prove equivalence of their maximizers.

Continuous-time recovery raises an additional stability issue. Even when the transition matrix and generator are uniquely determined at the population level, a small eigenvalue of $\mathbf{I}_K - \mathbf{C}_\pi$ makes the logarithmic inversion sensitive to covariance error. Estimated covariances can also yield matrices outside the stochastic-matrix or generator constraints. Quantifying this sensitivity and enforcing feasibility are natural next steps for a statistical estimator based on the inversion.

The observation model assumes independent units with a common time-homogeneous generator and exact histogram observations. The conditional-covariance inversion additionally assumes reversibility and an attainable equilibrium source. Applying it to one naturally evolving trajectory requires a method for estimating the relevant conditional covariance from varying source compositions. Stationary lagged moments offer an existing alternative for identifying $\mathbf{P}$ in that setting, including nonreversible chains; recovery of a nonreversible generator remains a separate embedding problem.

The Gaussian approximation has its clearest justification for fixed state dimension and strictly positive transition probabilities, with relative probability or log-likelihood accuracy in the central count region. This does not establish approximation of scores or Fisher information, uniform accuracy over an unrestricted optimization domain, or validity when the number of states grows with population size. The larger-state and alternative-topology experiments provide evidence within the configurations examined. Further work could develop parameter-uniform approximation bounds, estimator asymptotics, and procedures accommodating observation error or heterogeneous unit dynamics.

7 Conclusion

Conditional equilibrium fluctuations retain information about hidden transition dynamics. For an irreducible reversible CTMC, the equilibrium composition and the destination covariance conditional on that composition determine both the finite-time transition matrix and its generator. This population-level inversion uses a different observable from stationary lagged-covariance recovery. The aggregate likelihood and its dynamic-programming evaluation support inference from finite histogram sequences. Sufficient identifiability conditions describe when known conditional laws determine labeled transition probabilities, while a Gaussian calculation characterizes the $N^{-1/2}$ information crossover for parameter changes preserving equilibrium. The numerical studies illustrate these results and the distinction between likelihood approximation and recovery of continuous-time rates. Statistical accuracy, numerical optimization, and stability of the inverse map remain separate from population-level identification.

A Derivation of the Aggregate Markov Property

Condition on the current histogram $\mathbf{Y}_{t-1} = \mathbf{y}$. The y_j units currently occupying source state j form a disjoint group. By the unit-level Markov property and conditional independence across units, their destination counts over $(\tau_{t-1}, \tau_t]$ depend on the earlier history only through the current source state and are distributed according to row j of $\mathbf{P}_t$. Hence the row-wise destination-count vectors are conditionally independent multinomials with laws given by (4). Summing these row vectors produces $\mathbf{Y}_t$, so its conditional law given the full aggregate history depends on that history only through $\mathbf{Y}_{t-1}$. Marginalizing over the compatible row-wise flows gives exactly the kernel in (7), which proves the aggregate Markov property.

B Proof of Theorem 1 and Identification Details

For a fixed transition matrix $\mathbf{P}$ and source histogram $\mathbf{y}$, write

$$L_j(\mathbf{s}) = \sum_{k=1}^K P_{jk} s_k, \quad G_{\mathbf{y}}(\mathbf{s}) = \prod_{j=1}^K L_j(\mathbf{s})^{y_j}.$$

We prove Theorem 1 using the multiplicity-profile argument and then record its single-source special case.

1. Multiple source compositions: for source histograms $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(M)}$, define

$$\mathbf{v}_j = (y_j^{(1)}, \dots, y_j^{(M)})^\top.$$

Because $\mathbf{P} = \exp\{\Delta\mathbf{Q}\}$ is nonsingular for finite Δ , its rows are pairwise distinct; hence the corresponding normalized linear factors L_j are pairwise distinct. Factor each observable pgf separately and normalize every recovered linear factor by $L(\mathbf{1}_K) = 1$, which removes the scalar ambiguity. The condition $\mathbf{v}_j \neq \mathbf{0}_M$ ensures that every factor L_j appears in at least one of the M pgfs. Because the normalized factors are distinct, occurrences of the same factor in different pgfs can be matched by equality of their coefficient vectors. For each recovered factor L , the observable multiplicity vector

$$(\text{mult}_{G_{\mathbf{y}^{(1)}}} L, \dots, \text{mult}_{G_{\mathbf{y}^{(M)}}} L)^\top$$

is therefore determined and equals $\mathbf{v}_j$ when $L = L_j$. Pairwise distinct profiles $\mathbf{v}_j \neq \mathbf{v}_\ell$ then assign each recovered factor to a unique source-state label. Thus the collection of conditional aggregate-transition laws identifies the labeled matrix $\mathbf{P}$.

2. One source composition: when $M = 1$, the profile condition becomes $y_j > 0$ for every j and $y_j \neq y_\ell$ for $j \neq \ell$. In this case,

$$G_{\mathbf{y}}(\mathbf{s}) = \prod_{j=1}^K L_j(\mathbf{s})^{y_j}.$$

The polynomial ring $\mathbb{R}[s_1, \dots, s_K]$ is a unique factorization domain and each nonconstant linear polynomial is irreducible. Thus the distinct factors are identified up to scale and permutation. The stochastic-row constraint removes the scale ambiguity because

$$L_j(\mathbf{1}_K) = \sum_{k=1}^K P_{jk} = 1,$$

and the pairwise-distinct multiplicities y_j remove the remaining permutation ambiguity. Hence every labeled row of $\mathbf{P}$ is recovered.

The profile condition is sufficient rather than necessary. Its role is to make explicit how variation in observable source compositions resolves the row-label ambiguity of the aggregate pgf. Within the continuous-time model, pairwise distinct rows require no additional assumption: $\mathbf{P} = \exp(\Delta \mathbf{Q})$ is nonsingular for every finite Δ , whereas a matrix with two identical rows is singular.

C Boundary behavior and existence of likelihood maximizers

First consider a constant observed sequence $\mathbf{Y}_0 = \dots = \mathbf{Y}_T = \mathbf{y}$ with every $y_j > 0$, and set $\boldsymbol{\pi} = \mathbf{y}/N$. To exhibit boundary behavior within the general generator space, choose any irreducible reversible generator $\mathbf{Q}_0$ having stationary distribution $\boldsymbol{\pi}$ and define

$$\mathbf{Q}_\varepsilon = \varepsilon \mathbf{Q}_0, \quad \mathbf{P}_{t,\varepsilon} = \exp\{\Delta_t \mathbf{Q}_\varepsilon\}, \quad \varepsilon > 0.$$

As $\varepsilon \downarrow 0$, each $\mathbf{P}_{t,\varepsilon}$ tends to the identity, so $\mathcal{K}_t(\mathbf{y}, \mathbf{y}; \mathbf{P}_{t,\varepsilon}) \rightarrow 1$ and $\ell_T(\mathbf{Q}_\varepsilon) \rightarrow 0$. Since every transition probability is at most one, $\ell_T(\mathbf{Q}) \leq 0$ for every admissible generator $\mathbf{Q}$; the constructed sequence therefore shows that the global supremum is zero. For every irreducible interior generator and every $\Delta_t > 0$, $\mathbf{P}_t$ is strictly positive, so a destination histogram different from $\mathbf{y}$ has positive probability, $\mathcal{K}_t(\mathbf{y}, \mathbf{y}; \mathbf{P}_t) < 1$, and hence $\ell_T(\mathbf{Q}) < 0$. Thus the supremum zero is not attained within the irreducible interior family. Over the full generator space $\mathcal{Q}$, however, $\mathbf{Q} = \mathbf{0}$ is admissible and gives $\mathbf{P}_t = \mathbf{I}_K$, so the exact likelihood attains its maximum at $\mathbf{Q} = \mathbf{0}$. This constant-sequence example therefore establishes nonattainment only for the irreducible interior family, not for the full generator space.

For the Gaussian criterion, stationarity implies $\boldsymbol{\pi} \mathbf{P}_{t,\varepsilon} = \boldsymbol{\pi}$, so every composition residual is zero. Because $\mathbf{Q}_0$ is reversible,

$$\mathbf{D}_\pi - \mathbf{P}_{t,\varepsilon}^\top \mathbf{D}_\pi \mathbf{P}_{t,\varepsilon} = -2\varepsilon \Delta_t \mathbf{D}_\pi \mathbf{Q}_0 + O(\varepsilon^2).$$

Here $\mathbf{D}_\pi = \text{diag}(\pi_1, \dots, \pi_K)$. The symmetric matrix $-\mathbf{D}_\pi \mathbf{Q}_0$ is a weighted graph Laplacian; for an irreducible chain, its reduced principal submatrix is positive definite. With $d = K - 1$, define

$$\mathbf{W}_t = -2\Delta_t (\mathbf{D}_\pi \mathbf{Q}_0)_{1:d, 1:d},$$

where the subscript $1:d, 1:d$ denotes the principal submatrix obtained by deleting the final row and column. Then $\mathbf{W}_t \succ 0$, and the reduced covariance has the form

$$\mathbf{V}_t(\mathbf{Q}_\varepsilon) = \varepsilon \mathbf{W}_t + O(\varepsilon^2), \quad \mathbf{W}_t \succ 0,$$

and $\log \det \mathbf{V}_t = d \log \varepsilon + O(1)$. It follows that

$$\tilde{\ell}_T(\mathbf{Q}_\varepsilon) = -\frac{Td}{2} \log \varepsilon + O(1) \longrightarrow +\infty.$$

The Gaussian objective is therefore unbounded above for this dataset, even along generators with positive-definite reduced conditional covariances.

A separate example establishes nonexistence of an exact maximum-likelihood estimator over the full generator space $\mathcal{Q}$. Take $N = 1$, $K = 2$, and $T = 1$, with observed histograms $\mathbf{Y}_0 = (1, 0)$ and $\mathbf{Y}_1 = (0, 1)$ separated by a fixed interval $\Delta > 0$. For $a > 0$, the admissible generator

$$\mathbf{Q}_a = \begin{pmatrix} -a & a \\ 0 & 0 \end{pmatrix}$$

has conditional likelihood

$$L_1(\mathbf{Q}_a \mid \mathbf{Y}_0) = [\exp(\Delta \mathbf{Q}_a)]_{12} = 1 - e^{-a\Delta} \longrightarrow 1 \quad \text{as } a \rightarrow \infty.$$

For any finite generator $\mathbf{Q} = (q_{jk}) \in \mathcal{Q}$, however, a unit initially in state 1 has probability $e^{\Delta q_{11}} > 0$ of making no jump during the interval. Consequently,

$$[\exp(\Delta \mathbf{Q})]_{12} \leq 1 - e^{\Delta q_{11}} < 1.$$

Thus the exact likelihood has supremum one (equivalently, the log-likelihood has supremum zero), but no finite generator attains it.

The two datasets establish distinct failures: the constant sequence makes the Gaussian criterion unbounded above, whereas the single observed state change yields a nonattained exact-likelihood supremum over $\mathcal{Q}$. Hence neither unrestricted argmax in Section 3 defines an estimator for every possible observed sequence.

D Exact computation and implementation details

The finite- N likelihood is a sum over contingency tables and is therefore combinatorial if evaluated naively. This section develops a dynamic-programming representation that performs the required convolution over source rows without enumerating all feasible full tables. The same forward-backward quantities also provide the conditional moments needed for latent-flow reconstruction and exact score evaluation.

D.1 Dynamic-programming formulation

The observed transition likelihood can be evaluated without enumerating all complete $K \times K$ contingency tables. For one interval, write

$$\mathbf{y} = \mathbf{Y}_{t-1}, \quad \mathbf{y}' = \mathbf{Y}_t, \quad \mathbf{P} = \mathbf{P}_t.$$

Write $K_{\mathbf{P}}(\mathbf{y}, \mathbf{y}') := \mathcal{K}_t(\mathbf{y}, \mathbf{y}'; \mathbf{P})$ for the conditional transition probability. Posterior formulas below apply when this probability is positive. For source bin j , let

$$\mathbf{x}_j = (X_{j1}, \dots, X_{jK}) \sim \text{Multinomial}(y_j, P_{j1}, \dots, P_{jK}),$$

independently across j . Then

$$\mathbf{y}' = \sum_{j=1}^K \mathbf{x}_j. \tag{55}$$

Define

$$\mathcal{X}_j = \left\{ \mathbf{x} \in \mathbb{N}_0^K : \sum_{k=1}^K x_k = y_j \right\} \quad (56)$$

and

$$g_j(\mathbf{x}) = \frac{y_j!}{\prod_{k=1}^K x_k!} \prod_{k=1}^K P_{jk}^{x_k}, \quad \mathbf{x} \in \mathcal{X}_j. \quad (57)$$

1. Forward recursion: for $r = 0, \dots, K$, define

$$A_r(\mathbf{z}) = \Pr \left(\sum_{j=1}^r \mathbf{x}_j = \mathbf{z} \right). \quad (58)$$

Initialize

$$A_0(\mathbf{0}) = 1, \quad A_0(\mathbf{z}) = 0 \quad \text{for } \mathbf{z} \neq \mathbf{0}.$$

Then

$$A_r(\mathbf{z}) = \sum_{\mathbf{x} \in \mathcal{X}_r} A_{r-1}(\mathbf{z} - \mathbf{x}) g_r(\mathbf{x}), \quad (59)$$

where terms with a negative component in $\mathbf{z} - \mathbf{x}$ are zero. Hence

$$K_{\mathbf{P}}(\mathbf{y}, \mathbf{y}') = A_K(\mathbf{y}'). \quad (60)$$

With the multi-index notation $\mathbf{s}^{\mathbf{x}} := \prod_{k=1}^K s_k^{x_k}$, the recursion is equivalent to sequential coefficient extraction because

$$\left(\sum_{k=1}^K P_{j,k} s_k \right)^{y_j} = \sum_{\mathbf{x} \in \mathcal{X}_j} g_j(\mathbf{x}) \mathbf{s}^{\mathbf{x}}, \quad (61)$$

so

$$\prod_{j=1}^r \left(\sum_{k=1}^K P_{j,k} s_k \right)^{y_j} = \sum_{\mathbf{z}} A_r(\mathbf{z}) \mathbf{s}^{\mathbf{z}}.$$

2. State reduction: at stage r ,

$$\sum_{k=1}^K z_k = \sum_{j=1}^r y_j,$$

so only $K - 1$ coordinates of $\mathbf{z}$ must be stored:

$$z_K = \sum_{j=1}^r y_j - \sum_{k=1}^{K-1} z_k. \quad (62)$$

For fixed target $\mathbf{y}'$, any state with $z_k > y'_k$ can be discarded.

3. Backward recursion: define

$$B_r(\mathbf{z}) = \Pr \left(\sum_{j=r}^K \mathbf{x}_j = \mathbf{z} \right), \quad r = 1, \dots, K+1. \quad (63)$$

Initialize

$$B_{K+1}(\mathbf{0}) = 1, \quad B_{K+1}(\mathbf{z}) = 0 \quad \text{for } \mathbf{z} \neq \mathbf{0},$$

and recurse via

$$B_r(\mathbf{z}) = \sum_{\mathbf{x} \in \mathcal{X}_r} g_r(\mathbf{x}) B_{r+1}(\mathbf{z} - \mathbf{x}). \quad (64)$$

Then

$$A_K(\mathbf{y}') = B_1(\mathbf{y}') = K_{\mathbf{P}}(\mathbf{y}, \mathbf{y}').$$

4. Posterior distribution of a source-row flow: for fixed j , define

$$C_{-j}(\mathbf{r}) = \Pr \left(\sum_{a \neq j} \mathbf{x}_a = \mathbf{r} \right). \quad (65)$$

Using the forward and backward messages,

$$C_{-j}(\mathbf{r}) = \sum_{\mathbf{z}} A_{j-1}(\mathbf{z}) B_{j+1}(\mathbf{r} - \mathbf{z}). \quad (66)$$

Therefore,

$$\Pr(\mathbf{x}_j = \mathbf{x} \mid \mathbf{y}, \mathbf{y}', \mathbf{P}) = \frac{g_j(\mathbf{x}) C_{-j}(\mathbf{y}' - \mathbf{x})}{K_{\mathbf{P}}(\mathbf{y}, \mathbf{y}')}, \quad \mathbf{x} \in \mathcal{X}_j. \quad (67)$$

5. Posterior expected transition counts: the posterior expected transition count for the $j \rightarrow k$ cell is

$$\bar{X}_{jk} = \frac{\sum_{\mathbf{x} \in \mathcal{X}_j} x_k g_j(\mathbf{x}) C_{-j}(\mathbf{y}' - \mathbf{x})}{K_{\mathbf{P}}(\mathbf{y}, \mathbf{y}')}. \quad (68)$$

This is the forward-backward form of the posterior mean used in the exact score (90).

If $y_j = 0$, the source row is identically zero and $\bar{X}_{jk} = 0$ for every k . For $y_j \geq 1$, using the multinomial identity

$$x_k g_j(\mathbf{x}) = y_j P_{jk} g_j^{(-)}(\mathbf{x} - \mathbf{e}_k),$$

where $\mathbf{e}_k$ is the k th coordinate vector in $\mathbb{R}^K$ and $g_j^{(-)}$ is the multinomial mass function with total count $y_j - 1$ and probabilities $P_{j\cdot}$, gives

$$\bar{X}_{jk} = \frac{y_j P_{jk}}{K_{\mathbf{P}}(\mathbf{y}, \mathbf{y}')} \sum_{\mathbf{u}} g_j^{(-)}(\mathbf{u}) C_{-j}(\mathbf{y}' - \mathbf{e}_k - \mathbf{u}). \quad (69)$$

Thus $y_j P_{jk}$ is the expected $j \rightarrow k$ flow before the destination histogram is observed, while the remaining factor adjusts this expectation to respect the observed destination margin.

6. Computational scaling: the recursion avoids explicit enumeration of all feasible $K \times K$ transition tables, but its state space still grows with the population size and state dimension. For a single aggregate transition with source histogram $\mathbf{y}$, define

$$n_j^{\text{F}} = \sum_{a=1}^j y_a, \quad n_j^{\text{B}} = \sum_{a=j}^K y_a,$$

and

$$S_j^{\text{F}} = \binom{n_j^{\text{F}} + K - 1}{K - 1}, \quad S_j^{\text{B}} = \binom{n_j^{\text{B}} + K - 1}{K - 1}, \quad R_j = |\mathcal{X}_j| = \binom{y_j + K - 1}{K - 1}.$$

Before target-specific pruning, S_j^{F} bounds the number of possible states of the forward message A_j , S_j^{B} bounds the number of possible states of the backward message B_j , and R_j is the number of row compositions convolved at source state j . Consequently, with $S_0^{\text{F}} = S_{K+1}^{\text{B}} = 1$, direct sparse evaluation of the forward and backward recursions requires at most

$$O \left[\sum_{j=1}^K R_j (S_{j-1}^{\text{F}} + S_{j+1}^{\text{B}}) \right] \quad (70)$$

state-composition updates.

For score evaluation, it is unnecessary to construct every value of the leave-one-out distribution C_{-j} . Equation (68) only requires $C_{-j}(\mathbf{y}' - \mathbf{x})$ for $\mathbf{x} \in \mathcal{X}_j$. For any such $\mathbf{x}$,

$$C_{-j}(\mathbf{y}' - \mathbf{x}) = \sum_{\mathbf{z}} A_{j-1}(\mathbf{z}) B_{j+1}(\mathbf{y}' - \mathbf{x} - \mathbf{z}).$$

With constant-time lookup of stored sparse-message entries, this contraction can be evaluated by iterating over whichever of A_{j-1} and B_{j+1} has smaller support. Thus all required leave-one-out values for row j have the unpruned bound

$$O[R_j \min\{S_{j-1}^F, S_{j+1}^B\}].$$

Once these values are available, accumulating the K posterior expected counts in that source row costs at most $O(KR_j)$. The complete forward-backward and posterior-flow calculation for one observed transition is therefore bounded by

$$O\left[\sum_{j=1}^K R_j (S_{j-1}^F + S_{j+1}^B + \min\{S_{j-1}^F, S_{j+1}^B\} + K)\right]. \quad (71)$$

The destination constraint $z_k \leq y'_k$ can reduce all three message-support terms substantially in practice, so (71) is an unpruned upper bound for the direct sparse recursion, not an assertion of tight or optimal complexity.

The largest possible unrestricted histogram state space is

$$\binom{N + K - 1}{K - 1}. \quad (72)$$

Hence, for fixed K , the message-state counts are polynomial in N ; the difficulty is that the polynomial degree itself increases with K , while the row-composition factors R_j also grow. This explains why exact dynamic programming is useful for small populations and small or moderate state spaces but becomes expensive under joint growth of N and K . The main-text expression (13) adds the costs of the matrix exponential and its parameter derivatives, repeated observed transitions, optimization evaluations, and multiple initializations.

Finally, sparsity of the transition graph applies to the instantaneous generator $\mathbf{Q}$, not generally to

$$\mathbf{P}(\Delta_t) = \exp(\Delta_t \mathbf{Q}).$$

A unit may make multiple jumps along the graph during one observation interval, so $\mathbf{P}_t$ can be dense even when $\mathbf{Q}$ is sparse. Restricting latent interval transition counts to edges of $\mathcal{E}$ would therefore not be consistent with the continuous-time model except as an infinitesimal-time approximation.

E Derivation of the Aggregate Central-Limit Approximation

Fix K , put $d = K - 1$, and fix a stochastic matrix $\mathbf{P}$. Consider source histograms $\mathbf{y}^{(N)} \in \mathcal{H}_{N,K}$ such that $\mathbf{y}^{(N)}/N \rightarrow \mathbf{p} \in \Delta_{K-1}$. Define

$$\mathbf{q}_j = (P_{j1}, \dots, P_{jd})^\top, \quad \mathbf{C}_j = \text{diag}(\mathbf{q}_j) - \mathbf{q}_j \mathbf{q}_j^\top, \quad \mathbf{V} = \sum_{j=1}^K p_j \mathbf{C}_j,$$

and assume $\mathbf{V} \succ 0$. For each unit ℓ , let $\mathbf{W}_\ell^{(N)} \in \{\mathbf{0}, \mathbf{e}_1, \dots, \mathbf{e}_d\}$ be its reduced destination indicator, where $\mathbf{e}_k$ is the k th coordinate vector in $\mathbb{R}^d$. A unit from source state j has mean $\mathbf{q}_j$ and covariance $\mathbf{C}_j$. Conditional on the source histogram these vectors are independent, and

$$\mathbf{Y}_-'^{(N)} = \sum_{\ell=1}^N \mathbf{W}_\ell^{(N)}, \quad \boldsymbol{\xi}_\ell^{(N)} = \mathbf{W}_\ell^{(N)} - \mathbb{E}[\mathbf{W}_\ell^{(N)}].$$

The centered summands are uniformly bounded. Thus, after division by $\sqrt{N}$, they satisfy the Lindeberg condition, and

$$\frac{1}{N} \sum_{\ell=1}^N \text{Cov}(\boldsymbol{\xi}_\ell^{(N)}) = \sum_{j=1}^K \frac{y_j^{(N)}}{N} \mathbf{C}_j \longrightarrow \mathbf{V}.$$

The multivariate Lindeberg-Feller theorem (or its scalar form applied through the Cramér-Wold device) gives

$$\frac{\mathbf{Y}_-'^{(N)} - (\mathbf{y}^{(N)} \mathbf{P})_-}{\sqrt{N}} \Rightarrow \mathcal{N}_d(\mathbf{0}, \mathbf{V}).$$

This is the distributional justification for (17). It does not itself approximate individual lattice probabilities; that requires the local argument below.

F Derivation of the Local Gaussian Approximation

The ordinary aggregate central-limit approximation describes the distribution of the centered and scaled histogram. The exact transition likelihood, however, is a lattice probability

$$K_{\mathbf{P}}(\mathbf{y}, \mathbf{y}') = \Pr(\mathbf{Y}'_- = \mathbf{y}'_- \mid \mathbf{y}),$$

not a continuous density. A local central limit theorem is therefore the appropriate justification for replacing the exact probability mass by a Gaussian expression.

Let

$$\hat{\mathbf{p}} = \frac{\mathbf{y}}{N}, \quad \hat{\mathbf{p}}' = \frac{\mathbf{y}'}{N}, \quad \mathbf{m}_N = N(\hat{\mathbf{p}} \mathbf{P})_-,$$

and define

$$\mathbf{V}_N = \sum_{j=1}^K \hat{p}_j \left[\text{diag}(\mathbf{q}_j) - \mathbf{q}_j \mathbf{q}_j^\top \right], \quad \boldsymbol{\Sigma}_N = N \mathbf{V}_N,$$

where $\mathbf{q}_j = (P_{j1}, \dots, P_{j,K-1})^\top$. Thus $\mathbf{m}_N$ and $\boldsymbol{\Sigma}_N$ are respectively the exact conditional mean and covariance of the reduced destination count vector.

We give sufficient conditions and a direct local-limit argument for the approximation used here. Fix K and a stochastic matrix $\mathbf{P}$ with $P_{jk} > 0$ for every j, k . These conditions hold for a fixed irreducible finite-state CTMC observed over a fixed positive interval. Source histograms may vary arbitrarily with N . Since each categorical law has support $\{\mathbf{0}, \mathbf{e}_1, \dots, \mathbf{e}_d\}$ with positive probabilities, each $\mathbf{C}_j = \text{diag}(\mathbf{q}_j) - \mathbf{q}_j \mathbf{q}_j^\top$ is positive definite. Consequently, the eigenvalues of $\mathbf{V}_N$ are bounded above and bounded away from zero uniformly over source histograms.

Let $\phi_{\mathbf{V}}(\mathbf{x})$ denote the density of $\mathcal{N}_d(\mathbf{0}, \mathbf{V})$. Then

$$\sup_{\mathbf{y} \in \mathcal{H}_{N,K}} \sup_{\mathbf{z} \in \mathbb{Z}^d} \left| N^{d/2} \Pr(\mathbf{Y}'_- = \mathbf{z} \mid \mathbf{y}) - \phi_{\mathbf{V}_N} \left(\frac{\mathbf{z} - \mathbf{m}_N}{\sqrt{N}} \right) \right| \longrightarrow 0.$$

To verify this, the characteristic function of a single reduced destination indicator from source j is

$$\psi_j(\mathbf{u}) = P_{jK} + \sum_{k=1}^d P_{jk} e^{i u_k}.$$

All support probabilities are positive, so $|\psi_j(\mathbf{u})| = 1$ on the torus $[-\pi, \pi]^d$ only at $\mathbf{u} = \mathbf{0}$. Outside any fixed neighborhood of zero, the maximum of these moduli over the finitely many rows is some $r < 1$; hence the characteristic function of the sum is bounded by r^N there. Near zero, the centered functions satisfy, uniformly in j ,

$$\log\left(e^{-i\mathbf{u}^\top \mathbf{q}_j} \psi_j(\mathbf{u})\right) = -\frac{1}{2}\mathbf{u}^\top \mathbf{C}_j \mathbf{u} + O(\|\mathbf{u}\|^3).$$

Thus the centered characteristic function χ_N of the sum obeys, on every bounded set of $\mathbf{v}$,

$$\chi_N(\mathbf{v}/\sqrt{N}) = \exp\left\{-\frac{1}{2}\mathbf{v}^\top \mathbf{V}_N \mathbf{v} + O(N^{-1/2}\|\mathbf{v}\|^3)\right\}.$$

Moreover, for sufficiently small fixed $\delta > 0$, positive definiteness gives a uniform bound $|\chi_N(\mathbf{v}/\sqrt{N})| \leq e^{-c\|\mathbf{v}\|^2}$ whenever $\|\mathbf{v}\| \leq \delta\sqrt{N}$, for some $c > 0$. Fourier inversion, with the change of variables $\mathbf{v} = \sqrt{N}\mathbf{u}$, now proves the displayed local limit: the bounded region converges by the expansion, the intermediate region is controlled by the Gaussian bound, and the remaining integral is at most a constant times $N^{d/2}r^N$. The Fourier phase has modulus one, making the error uniform in $\mathbf{z}$; all bounds are also uniform in the source histogram.

In particular, uniformly for targets satisfying $\|\mathbf{y}'_- - \mathbf{m}_N\| \leq M\sqrt{N}$ for any fixed $M < \infty$,

$$K_{\mathbf{P}}(\mathbf{y}, \mathbf{y}') = \frac{\exp\{-\frac{1}{2}(\mathbf{y}'_- - \mathbf{m}_N)^\top \Sigma_N^{-1}(\mathbf{y}'_- - \mathbf{m}_N)\}}{(2\pi)^{d/2} \det(\Sigma_N)^{1/2}} + o(N^{-d/2}). \quad (73)$$

On this central region the Gaussian term is bounded below by a positive constant times $N^{-d/2}$, so the relative error is $o(1)$ and the log-probability error is $o(1)$. This establishes the transition approximation for fixed positive $\mathbf{P}$. It does not establish uniform approximation over the unrestricted generator space, convergence of likelihood maximizers, convergence of score derivatives or Fisher information, or a growing- K result. For fixed T , the log-probability errors can be summed when each transition satisfies the stated central-region condition; a growing observation horizon requires additional control.

Writing

$$\mathbf{r}_N = \hat{\mathbf{p}}'_- - (\hat{\mathbf{p}}\mathbf{P})_-, \quad (74)$$

and using $\Sigma_N = N\mathbf{V}_N$, the approximation becomes

$$K_{\mathbf{P}}(\mathbf{y}, \mathbf{y}') \approx \frac{\exp\{-\frac{N}{2}\mathbf{r}_N^\top \mathbf{V}_N^{-1} \mathbf{r}_N\}}{(2\pi N)^{d/2} \det(\mathbf{V}_N)^{1/2}}. \quad (75)$$

For interval t and generator $\mathbf{Q}$, let

$$\begin{aligned} \mathbf{P}_t(\mathbf{Q}) &= \exp\{\Delta_t \mathbf{Q}\}, \\ \mathbf{r}_t(\mathbf{Q}) &= \hat{\mathbf{p}}_{t,-} - (\hat{\mathbf{p}}_{t-1} \mathbf{P}_t(\mathbf{Q}))_-, \end{aligned} \quad (76)$$

and

$$\mathbf{V}_t(\mathbf{Q}) = \sum_{j=1}^K \hat{p}_{t-1,j} \left[\text{diag}(\mathbf{q}_{t,j}) - \mathbf{q}_{t,j} \mathbf{q}_{t,j}^\top \right], \quad (77)$$

where

$$\mathbf{q}_{t,j} := (P_{j1}^{(t)}, \dots, P_{j,K-1}^{(t)})^\top$$

is the column vector containing the first $K-1$ entries of row j of $\mathbf{P}_t(\mathbf{Q})$. Up to constants independent of $\mathbf{Q}$, the resulting large- N log-likelihood is

$$\tilde{\ell}_T(\mathbf{Q}) = -\frac{1}{2} \sum_{t=1}^T \left[\log \det \mathbf{V}_t(\mathbf{Q}) + N \mathbf{r}_t(\mathbf{Q})^\top \mathbf{V}_t(\mathbf{Q})^{-1} \mathbf{r}_t(\mathbf{Q}) \right]. \quad (78)$$

When the maximum is attained, this yields the approximate estimator

$$\hat{\mathbf{Q}}_G \in \arg \max_{\mathbf{Q} \in \mathcal{Q}_G(\mathcal{Y}_T)} \tilde{\ell}_T(\mathbf{Q}). \quad (79)$$

The existence qualification and the compact constrained alternative are discussed in Section 3.4; the local approximation alone does not imply existence of a maximizer.

Equation (19) is the main-text form of this approximation. It is not merely least squares applied to the deterministic mean relation. The transition mechanism enters both the mean and the covariance. Consequently, both the Mahalanobis mismatch term and the log-determinant term in (78) carry information about $\mathbf{Q}$.

G Proof of Theorem 2

Equation (29) gives $\mathbf{S}^2 = \mathbf{I}_K - \mathbf{C}_\pi$. Because

$$\mathbf{H} = \mathbf{D}_\pi^{1/2} \mathbf{Q} \mathbf{D}_\pi^{-1/2}$$

is symmetric under reversibility, $\mathbf{S} = e^{\Delta \mathbf{H}}$ is symmetric positive definite. Hence $\mathbf{S}$ is the unique principal symmetric positive-definite square root of $\mathbf{I}_K - \mathbf{C}_\pi$, which yields (31).

Since $\mathbf{S} \succ 0$, write its spectral decomposition as $\mathbf{S} = \mathbf{U} \text{diag}(\lambda_1, \dots, \lambda_K) \mathbf{U}^\top$ with $\lambda_i > 0$. Spectral calculus gives the unique real symmetric logarithm

$$\log \mathbf{S} = \mathbf{U} \text{diag}(\log \lambda_1, \dots, \log \lambda_K) \mathbf{U}^\top.$$

Because $\mathbf{H}$ is real symmetric and $\mathbf{S} = e^{\Delta \mathbf{H}}$, it follows that $\Delta \mathbf{H} = \log \mathbf{S}$. Therefore

$$\mathbf{H} = \Delta^{-1} \log \mathbf{S} = \frac{1}{2\Delta} \log(\mathbf{I}_K - \mathbf{C}_\pi).$$

Transforming back gives (32), completing the recovery of $\mathbf{P}$ and $\mathbf{Q}$.

H Proof of Theorem 3

Because $\boldsymbol{\pi} \mathbf{P}(\boldsymbol{\eta}) = \boldsymbol{\pi}$ for every admissible $\boldsymbol{\eta}$, differentiation with respect to η_r gives $\boldsymbol{\pi} \mathbf{P}_r = \mathbf{0}_K^\top$. Hence (37) holds. Along the local sequence (38),

$$\mathbf{m}_r(\boldsymbol{\rho}_N, \boldsymbol{\eta}) = N^{-\alpha} [\mathbf{h} \mathbf{P}_r]_- + o(N^{-\alpha}).$$

Since $\alpha > 0$, $\boldsymbol{\rho}_N \rightarrow \boldsymbol{\pi}$. Continuity and positive definiteness therefore give

$$\mathbf{V}_{\boldsymbol{\rho}_N}^{-1} = \mathbf{V}_{\boldsymbol{\pi}}^{-1} + o(1), \quad \mathbf{V}_r(\boldsymbol{\rho}_N) \rightarrow \mathbf{V}_{\boldsymbol{\pi}, r}.$$

Substituting these expansions into (36) yields

$$N \mathbf{m}_r^\top \mathbf{V}_{\boldsymbol{\rho}_N}^{-1} \mathbf{m}_s = N^{1-2\alpha} A_{rs}(\mathbf{h}) + o(N^{1-2\alpha})$$

and

$$\frac{1}{2} \text{tr}(\mathbf{V}_{\boldsymbol{\rho}_N}^{-1} \mathbf{V}_r \mathbf{V}_{\boldsymbol{\rho}_N}^{-1} \mathbf{V}_s) = B_{rs} + o(1).$$

Combining the two terms gives (39), with the two remainder terms retained separately to display the convergence of the fluctuation contribution to B_{rs} . The transition between the two contributions occurs when $N^{1-2\alpha}$ is of constant order, equivalently when $\|\boldsymbol{\rho}_N - \boldsymbol{\pi}\|$ is of order $N^{-1/2}$.

I Reversible mobility parameterization used in the numerical specialization

For a reversible generator with stationary distribution $\boldsymbol{\pi}$, detailed balance on any allowed edge $\{j, k\} \in \mathcal{E}$ requires

$$\pi_j q_{jk} = \pi_k q_{kj}.$$

Thus each edge can be written in terms of a positive equilibrium edge flow c_{jk} as

$$q_{jk} = \frac{c_{jk}}{\pi_j}, \quad q_{kj} = \frac{c_{jk}}{\pi_k}.$$

Writing

$$c_{jk} = \mu_{jk} \sqrt{\pi_j \pi_k}$$

gives

$$q_{jk} = \mu_{jk} \sqrt{\frac{\pi_k}{\pi_j}}, \quad q_{kj} = \mu_{jk} \sqrt{\frac{\pi_j}{\pi_k}}. \quad (80)$$

This separates the equilibrium weighting from the symmetric coupling strength. Indeed, with

$$\mathbf{H} = \mathbf{D}_\pi^{1/2} \mathbf{Q} \mathbf{D}_\pi^{-1/2}, \quad \mathbf{D}_\pi = \text{diag}(\boldsymbol{\pi}),$$

detailed balance implies that $\mathbf{H}$ is symmetric, and for every allowed edge

$$[\mathbf{H}]_{jk} = \sqrt{\pi_j} q_{jk} \pi_k^{-1/2} = \mu_{jk}.$$

Hence μ_{jk} is not itself either directional rate; it is the equilibrium-normalized symmetric coupling associated with edge $\{j, k\}$. Because generator entries have units of inverse time and the equilibrium probabilities are dimensionless, μ_{jk} also has units of inverse time. The nearest-neighbor construction is obtained by taking $\mathcal{E}$ to be the path graph and writing $\mu_j = \mu_{j,j+1}$.

J Gradient details for the reversible nearest-neighbor numerical specialization

For the numerical studies of Section 5, we optimize over the reversible nearest-neighbor parameterization $(\boldsymbol{\pi}, \boldsymbol{\mu})$. This appendix records the corresponding score, parameterization, and derivative formulas. These formulas are specific to that numerical specialization and are not assumptions of the general likelihood in Section 3. The Fréchet derivative of the matrix exponential may be evaluated directly or through the standard block-exponential identity

$$\exp \begin{pmatrix} \mathbf{A} & \mathbf{E} \\ \mathbf{0} & \mathbf{A} \end{pmatrix} = \begin{pmatrix} \exp(\mathbf{A}) & L_{\exp}(\mathbf{A}, \mathbf{E}) \\ \mathbf{0} & \exp(\mathbf{A}) \end{pmatrix}. \quad (81)$$

For the log-mobility coordinate β_r , $\mu_r = \exp(\beta_r)$ and

$$\frac{\partial q_{r,r+1}}{\partial \beta_r} = q_{r,r+1}, \quad \frac{\partial q_{r+1,r}}{\partial \beta_r} = q_{r+1,r}, \quad (82)$$

with all other off-diagonal derivatives equal to zero; diagonal derivatives enforce zero row sums.

For the baseline-logit coordinates,

$$\frac{\partial \log \pi_j}{\partial \alpha_r} = \mathbf{1}\{j = r\} - \pi_r, \quad j = 1, \dots, K, \quad (83)$$

where $\mathbf{1}\{K = r\} = 0$. Since each neighboring rate depends only on a ratio of adjacent stationary probabilities, the common softmax normalization cancels and

$$\frac{\partial q_{j,j+1}}{\partial \alpha_r} = \frac{q_{j,j+1}}{2} (\mathbf{1}\{j+1 = r\} - \mathbf{1}\{j = r\}), \quad (84)$$

$$\frac{\partial q_{j+1,j}}{\partial \alpha_r} = \frac{q_{j+1,j}}{2} (\mathbf{1}\{j = r\} - \mathbf{1}\{j+1 = r\}). \quad (85)$$

Again, the diagonal derivatives are the negatives of the corresponding off-diagonal row sums. These sparse expressions avoid numerical differentiation of the generator and are the matrices $\mathbf{Q}_r$ used in (91).

J.1 Exact observed-data score and unconstrained optimization

For numerical maximization, let $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{K-1}$ and define

$$\pi_j(\boldsymbol{\alpha}) = \frac{\exp(\alpha_j)}{1 + \sum_{\ell=1}^{K-1} \exp(\alpha_\ell)}, \quad j = 1, \dots, K-1, \quad (86)$$

$$\pi_K(\boldsymbol{\alpha}) = \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(\alpha_\ell)}, \quad (87)$$

$$\mu_j(\boldsymbol{\beta}) = \exp(\beta_j), \quad j = 1, \dots, K-1. \quad (88)$$

With $\boldsymbol{\vartheta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$, define the reparameterized exact log-likelihood

$$\ell_T^u(\boldsymbol{\vartheta}) := \ell_T(\boldsymbol{\pi}(\boldsymbol{\alpha}), \boldsymbol{\mu}(\boldsymbol{\beta})).$$

Define $\Theta := \text{int}(\Delta_{K-1}) \times (0, \infty)^{K-1}$, the parameter space of $(\boldsymbol{\pi}, \boldsymbol{\mu})$ for the nearest-neighbor family. This change of coordinates is a bijection between $\mathbb{R}^{2K-2}$ and Θ . Hence global maximizers correspond whenever they exist. The transformation enforces positivity and normalization, but does not prevent maximizing sequences from escaping to infinity in the transformed coordinates. It imposes no additional lower or upper bounds on the original parameters.

For the irreducible nearest-neighbor specialization considered here and $\Delta_t > 0$, every entry of $\mathbf{P}_t = \exp(\Delta_t \mathbf{Q})$ is strictly positive, so the log-transition derivatives below are well defined.

Writing the complete-data contribution for interval t as

$$\ell_{c,t}(\boldsymbol{\vartheta}; \mathbf{X}_t) = \sum_{j=1}^K \sum_{k=1}^K X_{jk}^{(t)} \log P_{jk}^{(t)}(\boldsymbol{\vartheta}) + C_t(\mathbf{X}_t), \quad (89)$$

Here $C_t(\mathbf{X}_t) = \sum_j \log(Y_{t-1,j}!) - \sum_{j,k} \log(X_{jk}^{(t)}!)$ is independent of the transition parameters. Fisher's identity gives the exact observed-data score

$$\nabla_{\boldsymbol{\vartheta}} \ell_T^u(\boldsymbol{\vartheta}) = \sum_{t=1}^T \sum_{j=1}^K \sum_{k=1}^K \bar{X}_{jk}^{(t)}(\boldsymbol{\vartheta}) \nabla_{\boldsymbol{\vartheta}} \log P_{jk}^{(t)}(\boldsymbol{\vartheta}). \quad (90)$$

For component ϑ_r , with $\mathbf{Q}_r = \partial \mathbf{Q} / \partial \vartheta_r$,

$$\frac{\partial \mathbf{P}_t}{\partial \vartheta_r} = L_{\exp}(\Delta_t \mathbf{Q}, \Delta_t \mathbf{Q}_r), \quad (91)$$

so that

$$\frac{\partial \ell_T^u}{\partial \vartheta_r} = \sum_{t,j,k} \bar{X}_{jk}^{(t)} \frac{[L_{\exp}(\Delta_t \mathbf{Q}, \Delta_t \mathbf{Q}_r)]_{jk}}{P_{jk}^{(t)}}. \quad (92)$$

J.2 Optimization details for the exact likelihood

For the numerical specialization of Section 5, the unconstrained coordinates $\boldsymbol{\vartheta}$ are mapped to $(\boldsymbol{\pi}, \boldsymbol{\mu})$ and hence to $\mathbf{Q}$. The exact-likelihood branch of Algorithm 1 evaluates $\mathcal{K}_t$ and $\bar{\mathbf{X}}_t$ using the forward-backward recursion above and evaluates the score using (92). For the irreducible nearest-neighbor specialization considered here and $\Delta_t > 0$, $P_{jk}^{(t)} > 0$ for all j, k , so the log-transition derivatives in (92) are well defined. The likelihood and score evaluations are exact up to numerical arithmetic.

For the population-size and observation-length comparisons, the supplied `fit_model.m` implementation uses MATLAB’s derivative-free `fminsearch`, rather than the quasi-Newton update in Algorithm 1. It sets `MaxFunEvals`=5000, `MaxIter`=2000, and both `TolX` and `TolFun` to 10^{-6} . The two methods use two matched starts: the first perturbs the true stationary weights and mobilities multiplicatively by $\exp(0.05Z)$ with independent standard normal Z , renormalizing the stationary weights; the second adds independent $\mathcal{N}(0, 0.25^2)$ perturbations in transformed coordinates. These starts use simulation truth and therefore do not assess initialization from data alone. For the Gaussian comparison, the implementation first attempts a Cholesky factorization of the symmetrized covariance. If it fails, it adds $\epsilon_t \mathbf{I}_d$, where $\epsilon_t = \max\{10^{-12}, 10^{-10} \text{tr}(\mathbf{V}_t)/d\}$, and retries. This numerical modification changes the criterion when activated and is separate from the unregularized theoretical estimator on $\mathcal{Q}_G(\mathcal{Y}_T)$. Invalid numerical likelihood evaluations receive infinite negative-log-likelihood cost. The routine selects the smallest returned objective across starts even when the associated exit flag is nonpositive; returned fits therefore need not be converged solutions. These settings describe those comparisons only, not every optimization routine used in Section 5. Reproducible reporting for other experiments should likewise specify stopping tolerances, numerical parameter restrictions, and unsuccessful-run handling. Fits with near-zero or very large mobilities should be assessed for sensitivity to these choices and for continued improvement along parameter sequences approaching the boundary. These diagnostics are distinct from the structural-identifiability analysis.

Statements and Declarations

Data availability. All numerical data reported in this article are generated synthetically from the stochastic models described in the manuscript.

Code availability. The MATLAB code used to generate the numerical results and figures in this article is available at <https://github.com/idonevat/Inferring-Transition-Dynamics-from-Histogram-T>

References

- F. Baltazar-Larios and L. J. R. Esparza. Statistical inference for partially observed markov-modulated diffusion risk model. *Methodology and Computing in Applied Probability*, 24(2): 571–593, 2022. doi: 10.1007/s11009-022-09932-7.
- Garrett Bernstein and Daniel Sheldon. Consistently estimating Markov chains with noisy aggregate data. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1142–1150. PMLR, 2016. URL <https://proceedings.mlr.press/v51/bernstein16.html>.
- P. J. Brown and C. D. Payne. Aggregate data, ecological regression, and voting transitions. *Journal of the American Statistical Association*, 81(394):452–460, 1986. doi: 10.1080/01621459.1986.10478290.

- C. E. Can, G. Ergun, and R. Soyer. Bayesian analysis of proportions via a hidden markov model. *Methodology and Computing in Applied Probability*, 24(4):3121–3139, 2022. doi: 10.1007/s11009-022-09971-0.
- M. Crowder and D. Stephens. On inference from markov chain macro-data using transforms. *Journal of Statistical Planning and Inference*, 141(9):3201–3216, 2011.
- B. A. Davis, C. R. Heathcote, and T. J. O’Neill. Estimating and interpolating a markov chain from aggregate data. *Biometrika*, 89(1):95–110, 2002. doi: 10.1093/biomet/89.1.95.
- D. L. Hawkins and C.-P. Han. Estimating transition probabilities from aggregate samples plus partial transition data. *Biometrics*, 56(3):848–854, 2000. doi: 10.1111/j.0006-341X.2000.00848.x.
- C. R. Heathcote and B. D. Puza. Estimating processes of counts from cross-sectional aggregate data, with an application to multistate life tables and health expectancies. *Australian & New Zealand Journal of Statistics*, 47(4):425–439, 2005. doi: 10.1111/j.1467-842X.2005.00406.x.
- C. Jia. A solution to the reversible embedding problem for finite markov chains. *Statistics & Probability Letters*, 116:122–130, 2016.
- J. D. Kalbfleisch and J. F. Lawless. Least-squares estimation of transition probabilities from aggregate data. *The Canadian Journal of Statistics*, 12(3):169–182, 1984.
- J. D. Kalbfleisch, J. F. Lawless, and W. M. Vollmer. Estimation in markov models from aggregate data. *Biometrics*, 39(4):907–919, 1983.
- Steven M. Kay. *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall, Englewood Cliffs, NJ, 1993.
- Gary King. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press, Princeton, NJ, 1997.
- J. F. Lawless and D. L. McLeish. The information in aggregate data from markov chains. *Biometrika*, 71(3):419–430, 1984.
- D. E. K. Martin and J. A. D. Aston. Distribution of statistics of hidden state sequences through the sum-product algorithm. *Methodology and Computing in Applied Probability*, 15(4):897–918, 2013. doi: 10.1007/s11009-012-9289-4.
- M. Mascherpa, A. Ringh, A. Taghvaei, and J. Karlsson. A convex approach for markov chain estimation from aggregate data via inverse optimal transport. *European Journal of Control*, page 101592, 2026. doi: 10.1016/j.ejcon.2026.101592.