Consistently Estimating Markov Chains with Noisy Aggregate Data

Garrett Bernstein   **Daniel Sheldon**

University of Massachusetts Amherst

Abstract

We address the problem of estimating the parameters of a time-homogeneous Markov chain given only noisy, aggregate data. This arises when a population of individuals behave independently according to a Markov chain, but individual sample paths cannot be observed due to limitations of the observation process or the need to protect privacy. Instead, only population-level counts of the number of individuals in each state at each time step are available. When these counts are exact, a conditional least squares (CLS) estimator is known to be consistent and asymptotically normal. We initiate the study of method of moments estimators for this problem to handle the more realistic case when observations are additionally corrupted by noise. We show that CLS can be interpreted as a simple "plug-in" method of moments estimator. However, when observations are noisy, it is not consistent because it fails to account for additional variance introduced by the noise. We develop a new, simpler method of moments estimator that bypasses this problem and is consistent under noisy observations.

1   Introduction

The problem of learning from aggregate data has arisen over the years in diverse fields including machine learning, statistics and econometrics, and social sciences. In each case, the goal is to make inferences or fit models at the level of individuals when the only available data is at the population level, for example, counts of the number of individuals with certain

properties. Example applications include: learning models of bird migration from citizen science count data (Sheldon et al., 2013; Liu et al., 2014), learning models of human mobility from data that is aggregated to maintain privacy (Sun et al., 2015), fitting models of voter turnout and demography from census data (King, 2013; Flaxman et al., 2015), and modeling of credit risk from historical data about the proportions of institutions with different credit ratings (Jones, 2005).

We consider the particular problem of estimating the parameters of a time-homogeneous Markov chain from noisy, aggregate data. In this problem, N individuals behave independently according to the same Markov chain for T time steps, and, at each time step, a noisy observation is made of the number of individuals in each state. The entire process is repeated K times, and the goal is to recover the transition probabilities of the Markov chain. We assume that the chains are started from the stationary distribution, so the entire process is stationary and the only parameters to estimate are the transition probabilities. The fundamental questions we seek to address are: Is it possible to recover transition probabilities given only aggregate data, and, if so, under what conditions? How much aggregate data is necessary to obtain accurate parameter estimates, and how does this compare to estimation with individual-level data?

This problem has previously arisen in two distinct settings. First, it has a long history in statistics and econometrics, where it is sometimes referred to as estimating Markov chains from "macro" data. The traditional approach is a conditional least squares (CLS) estimator (Miller, 1952; Madansky, 1959; Lee et al., 1970; Aigner and Goldfeld, 1974; Van Der Plas, 1983; Kalbfleisch et al., 1983). When observations are exact, it is known that the CLS estimator is consistent and asymptotically normal as $T \to \infty$ (Van Der Plas, 1983), which answers one of our questions: it is indeed possible to learn transition probabilities with only aggregate data. However, little is known about CLS or other estimators for the case when observations are noisy. It was previously presumed that CLS is not

consistent when observations follow a simple binomial noise model (MacRae, 1977). Our analysis highlights exactly why this is true and suggests alternate estimators that are consistent.

The second setting where our problem has appeared is in the context of *collective graphical models* (CGMs) (Sheldon and Dietterich, 2011), a recent formalism for inference and learning with aggregate data. In CGMs, individual-level data are generated by any graphical model, and observations are made of contingency tables (counts of the number of times different variable configurations appear in the population). The model we consider here is the special case of CGMs where the individual model is a time-homogeneous Markov chain. Unlike the prior work on aggregate Markov chains, CGMs explicitly model noise in the observations (Sheldon and Dietterich, 2011). However, the CGM literature has focused primarily on inference, and uses expectation maximization for learning, which is effective in certain cases but provides no guarantees (Sheldon and Dietterich, 2011; Sheldon et al., 2013; Liu et al., 2014; Sun et al., 2015). Our work contributes the first learning method with guarantees of any kind for a subclass of CGMs.

In this paper, we initiate the study of method of moments estimators for aggregate Markov chains to explicitly deal with imperfect observations. This is an important practical issue: aggregate data are rarely complete surveys of a population, so they should be considered *noisy* counts of the number of individuals in each state. The method of moments viewpoint yields a number of useful observations. First, we show that CLS can be interpreted as a simple "plug-in" method of moments estimator. However, when observations are noisy, it is not consistent because it fails to account for additional variance introduced by the noise. Second, we develop a new and simpler method of moments estimator that bypasses the issue faced by CLS and is consistent with noisy observations.

Our primary contribution is to develop the first estimator with comprehensive theoretical guarantees for estimation of Markov chains from noisy, aggregate data. We show that our new method of moments estimator is consistent in both the *time-average* ($T \to \infty$, fixed K) and *ensemble-average* ($K \to \infty$, fixed T) settings with observations from a broad class of noise models. We show through both theoretical and empirical results that the squared error of our estimator decays as $\mathcal{O}(1/TK)$. One previous work (MacRae, 1977) considered the problem of estimating Markov chains with aggregate data corrupted by binomial noise. Based on the presumption that CLS was not consistent, which we confirm here, MacRae proposed a "limited information" estimator. However, that estimator relies on

signal in the time-varying marginals of the process, and we show that it is *not* consistent for the stationary process we consider here.

In the remainder of the paper, we describe the model and problem statement (Section 2), introduce method of moments estimators, including CLS, for perfect observations (Section 3), develop our new estimator for noisy observations (Section 4), prove theoretical guarantees (Section 5), and then evaluate various estimators empirically (Section 6).

2 Model and Problem Statement

Our model for individuals is an ergodic, time-homogeneous Markov chain on state space $\{1, \ldots, S\}$. The probability for the state trajectory $x_1, \ldots, x_T$ is:

$$p(x_1, \ldots, x_T) = \pi(x_1) \prod_{t=1}^{T-1} P(x_t, x_{t+1}),$$

where P is the $S \times S$ transition matrix, whose entries $P(i, j)$ specify the probability of transitioning from state i to state j, and $\boldsymbol{\pi}$ is the stationary distribution, i.e., the unique vector $\boldsymbol{\pi}$ such that $\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$.

Aggregate data is generated by N individuals independently transitioning from state to state according to the same Markov chain. Let $x_t^{(m)}$ be the state of the mth individual at time t, and let $\mathbf{n}_t$ be the vector with entries $n_t(i) = \sum_{m=1}^N [x_t^{(m)} = i]$ that count the number of individuals in each state at time t (here $[\cdot]$ denotes an indicator function). The vectors $\mathbf{n}_1, \ldots, \mathbf{n}_T$ constitute the aggregate data. We further assume that observations are noisy, so the observed data are vectors $\mathbf{y}_1, \ldots, \mathbf{y}_T$ that depend probabilistically on the aggregate data through a noise model $p(\mathbf{y}_t \mid \mathbf{n}_t)$. For example, a model we consider later is one where individuals are observed with probability α, so $y_t(i) \sim \text{Binomial}(n_t(i), \alpha)$. The entire aggregate process is repeated K times, independently, to yield data vectors $\mathbf{y}_t^{(k)}$ for $k \in \{1, \ldots, K\}$ and $t \in \{1, \ldots, T\}$. The goal is to estimate P from the noisy, aggregate data $\{\mathbf{y}_t^{(k)}\}$. The plate diagram for one realization of the aggregate process is shown in Figure 1.

Stationarity. We assume that chains start in the stationary distribution for simplicity and because this is the most difficult setting for estimation. A main focus of our work is the case when $T \to \infty$, in which case our estimators and asymptotic guarantees apply without modification for arbitrary initial distributions; we simply need to wait slightly longer for the chains to mix. It is also easy to modify our estimators to explicitly model the non-stationary initial distributions;
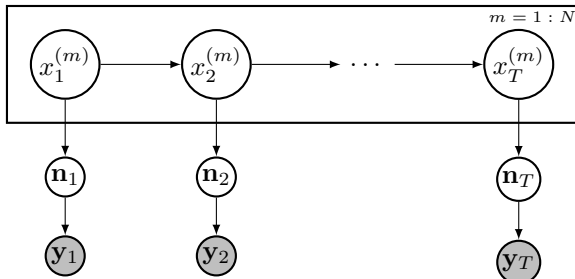
Figure 1: Plate Model

we describe this in Section 4.3. To see why estimation is most difficult in the stationary setting, and to set up our later analysis, note that because each chain is started in the stationary distribution and is time-homogeneous, the process is (strongly) stationary: the joint distribution of $(x_{t_1}, \ldots, x_{t_k})$ is the same as that of $(x_{t_1+a}, \ldots, x_{t_k+a})$ for any subset of times $t_1, \ldots, t_k$ and any time lag a. In particular, the marginal distribution of x_t is equal to $\boldsymbol{\pi}$ for all t. This means that, marginally, each vector $\mathbf{n}_t$ is a multinomial draw from the stationary distribution; it is quite clear that we can estimate $\boldsymbol{\pi}$ accurately from these vectors, but much less obvious that they contain enough information to estimate P.

3 Method of Moments

We will now describe how the method of moments applies to this problem by showing that the first and second moments of the aggregate process, which can be estimated from data, uniquely identify the marginals of the process, which in turn uniquely identify the transition matrix.

Marginals. Let $\boldsymbol{\mu}_t \in \mathbb{R}^S$ and $\boldsymbol{\mu}_{t,t+1} \in \mathbb{R}^{S \times S}$ be the vector and matrix, respectively, of *single* and *pairwise* marginals of the Markov process, defined by:

$$\boldsymbol{\mu}_t(i) = \Pr(x_t = i)$$
$$\boldsymbol{\mu}_{t,t+1}(i,j) = \Pr(x_t = i, x_{t+1} = j).$$

Since our process is stationary, the marginals $\boldsymbol{\mu}_t$ and $\boldsymbol{\mu}_{t,t+1}$ do not depend on t and $\boldsymbol{\mu}_t = \boldsymbol{\pi}$ for all t. However, we retain the dependence on t to make it more clear how the results generalize to non-stationary Markov chains.

It is clear that one can recover the transition matrix P from the pairwise marginals, since $P(i,j) = \Pr(x_{t+1} = j | x_t = i) = \mu_{t,t+1}(i,j)/\mu_t(i)$, or, in matrix form:

$$P = \text{diag}(\boldsymbol{\mu}_t)^{-1} \boldsymbol{\mu}_{t,t+1}.$$

Thus, if we can consistently estimate the marginals, we can recover P.

Moments. We will see that the marginals are encoded in a simple way in the first and second moments of the *aggregate* Markov process $\mathbf{n}_1, \ldots, \mathbf{n}_T$. Denote the mean vector at each time step (the first moments) by $\mathbf{m}_t = \mathbb{E}[\mathbf{n}_t]$, and denote the covariance matrices at time lags zero and one, respectively, by $\Sigma_t := \text{Var}(\mathbf{n}_t)$ and $\Sigma_{t,t+1} := \text{Cov}(\mathbf{n}_t, \mathbf{n}_{t+1})$. Denote the non-central second moments at time lags zero and one, respectively, by $\Lambda_t := \mathbb{E}[\mathbf{n}_t \mathbf{n}_t^T]$ and $\Lambda_{t,t+1} := \mathbb{E}[\mathbf{n}_t \mathbf{n}_{t+1}^T]$. Each of these quantities is estimable from the aggregate data. The following proposition shows that they also encode the single and pairwise marginals of the process in a straightforward way:

Proposition 1. *For all t, the following are true:*

$$\mathbf{m}_t = N\boldsymbol{\mu}_t \tag{1}$$
$$\Sigma_t = N\big(\text{diag}(\boldsymbol{\mu}_t) - \boldsymbol{\mu}_t \boldsymbol{\mu}_t^T\big) \tag{2}$$
$$\Sigma_{t,t+1} = N\big(\boldsymbol{\mu}_{t,t+1} - \boldsymbol{\mu}_t \boldsymbol{\mu}_{t+1}^T\big) \tag{3}$$
$$\Lambda_t = N\big(\text{diag}(\boldsymbol{\mu}_t) + (N-1)\boldsymbol{\mu}_t \boldsymbol{\mu}_t^T\big) \tag{4}$$
$$\Lambda_{t,t+1} = N\big(\boldsymbol{\mu}_{t,t+1} + (N-1)\boldsymbol{\mu}_t \boldsymbol{\mu}_{t+1}^T\big). \tag{5}$$

Proof. The marginal distribution of $\mathbf{n}_t$ is Multinomial$(N, \boldsymbol{\mu}_t)$. Equations (1) and (2) are well known formulas for the mean vector and covariance matrix of this multinomial distribution. Equation (3) was proved by Liu et al. (2014). Equations (4) and (5) follow from the previous equations and the fact that $\mathbb{E}[\mathbf{n}_s \mathbf{n}_t^T] = \text{Cov}(\mathbf{n}_s, \mathbf{n}_t^T) + \mathbb{E}[\mathbf{n}_s]\mathbb{E}[\mathbf{n}_t^T]$. $\qquad\square$

3.1 Conditional Least Squares

CLS (Miller, 1952) is based on the observation that, for all t, we expect $\mathbf{n}_t$ to be close to its conditional expectation given $\mathbf{n}_{t-1}$, which is $\mathbf{n}_{t-1}^T P$. In other words, we expect the following for all t:

$$\mathbf{n}_{t-1}^T P \approx \mathbf{n}_t.$$

We collect this into a least squares system by letting $X = \begin{bmatrix} \mathbf{n}_1 \ \mathbf{n}_2 \ \ldots \ \mathbf{n}_{T-1} \end{bmatrix}^T$ and $Y = \begin{bmatrix} \mathbf{n}_2 \ \mathbf{n}_3 \ \ldots \ \mathbf{n}_T \end{bmatrix}^T$. We can then write the least squares problem as:

$$\hat{P}_{\text{CLS}} = \text{argmin}_P \|XP - Y\|_F^2.$$

where $\|\cdot\|_F^2$ is the squared Frobenius norm, which is the sum of squares of the matrix entries. It is well known that the solution is given by

$$\hat{P}_{\text{CLS}} = (X^T X)^{-1} X^T Y.$$

Interpretation as Method of Moments We first note that $X^T X$ and $X^T Y$ are proportional to the straightforward empirical estimators of the non-central

moments Λ_t and $\Lambda_{t,t+1}$:

$$(X^T X)_{ij} = \sum_{t=1}^{T-1} n_t(i) n_t(j)$$

$$(X^T Y)_{ij} = \sum_{t=1}^{T-1} n_t(i) n_{t+1}(j).$$

To emphasize this point, we write $\hat{\Lambda}_t = (T-1)^{-1} X^T X$ and $\hat{\Lambda}_{t,t+1} = (T-1)^{-1} X^T Y$, and can then equivalently write the CLS estimator as

$$\hat{P}_{\text{CLS}} = \hat{\Lambda}_t^{-1} \hat{\Lambda}_{t,t+1}.$$

The following proposition shows that this is, in fact, a simple "plug-in" estimator for P using the empirical estimates of the non-central moments in place of their true counterparts.

Proposition 2. $P = \Lambda_t^{-1} \Lambda_{t,t+1}$.

Proof. It is straightforward to show that Λ_t is invertible (see supplementary material). We then verify the equivalent statement that $\Lambda_t P = \Lambda_{t,t+1}$:

$$\begin{aligned} \Lambda_t P &= N\big(\operatorname{diag}(\boldsymbol{\mu}_t) + (N-1)\boldsymbol{\mu}_t \boldsymbol{\mu}_t^T\big) P \\ &= N\big(\operatorname{diag}(\boldsymbol{\mu}_t) P + (N-1)\boldsymbol{\mu}_t \boldsymbol{\mu}_t^T P\big) \\ &= N\big(\boldsymbol{\mu}_{t,t+1} + (N-1)\boldsymbol{\mu}_t \boldsymbol{\mu}_{t+1}^T\big). \end{aligned}$$

This matches the definition of $\Lambda_{t,t+1}$ (Equation 5), as desired. $\qquad\square$

3.2 A Simpler Estimator

Proposition 1 (Equation 3) suggests a simpler way to recover the pairwise marginal $\boldsymbol{\mu}_{t,t+1}$, and hence P, from the moments of the aggregate process:

$$\boldsymbol{\mu}_{t,t+1} = N^{-1} \Sigma_{t,t+1} + \boldsymbol{\mu}_t \boldsymbol{\mu}_{t+1}^T.$$

So, a simple method of moments estimator is obtained by replacing the moments by their empirical estimators:

$$\hat{P}_{\text{MoM}} = \operatorname{diag}(\hat{\boldsymbol{\mu}}_t)^{-1}\big(N^{-1} \hat{\Sigma}_{t,t+1} + \hat{\boldsymbol{\mu}}_t \hat{\boldsymbol{\mu}}_{t+1}^T\big). \qquad (6)$$

We will describe the estimator in more detail (see Algorithm 1) after introducing noisy observations in the next section. We will observe that this simpler estimator performs better asymptotically than CLS for noisy observations.

4 Noisy Observations

So far we have assumed that the exact aggregate data vectors $\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_T$ are observed and thus we can estimate the moments. What happens if we only observe noisy vectors $\mathbf{y}_1, \ldots, \mathbf{y}_T$? It is not hard to imagine that for simple enough noise models, it will still be possible to consistently estimate the moments of $\mathbf{n}$; in particular, this is always possible if there is a bijection between the moments of $\mathbf{n}$ and $\mathbf{y}$. Our approach will be to define a large class of noise models for which it is straightforward to recover certain moments of $\mathbf{n}$ from those of $\mathbf{y}$. In the process of setting up this connection we will see exactly why CLS is *not* consistent when applied to the noisy vectors $\mathbf{y}_1, \ldots, \mathbf{y}_T$.

Example: Inconsistency of CLS. We will start with a very simple example to illustrate the main points. Suppose that $\mathbf{y}_t = \mathbf{n}_t + \epsilon_t$ where ϵ_t is a zero-mean noise vector that is independent of all other random variables. Then, one can easily verify the following relationships:

$$\begin{aligned} \mathbb{E}\left[\mathbf{y}_t\right] &= \mathbb{E}\left[\mathbf{n}_t\right] \\ \operatorname{Var}\left(\mathbf{y}_t\right) &= \operatorname{Var}\left(\mathbf{n}_t\right) + \operatorname{Var}\left(\epsilon_t\right) \\ \mathbb{E}\left[\mathbf{y}_t \mathbf{y}_t^T\right] &= \mathbb{E}\left[\mathbf{n}_t \mathbf{n}_t^T\right] + \operatorname{Var}\left(\epsilon_t\right) \\ \operatorname{Cov}(\mathbf{y}_t, \mathbf{y}_{t+1}) &= \operatorname{Cov}(\mathbf{n}_t, \mathbf{n}_{t+1}) \\ \mathbb{E}\left[\mathbf{y}_t \mathbf{y}_{t+1}^T\right] &= \mathbb{E}\left[\mathbf{n}_t \mathbf{n}_{t+1}^T\right]. \end{aligned}$$

In particular, if we just pretend the data is not noisy and use $\mathbf{y}$ in our estimators in place of $\mathbf{n}$, this *almost* does the right thing. It is only the zero-lag moments $\operatorname{Var}\left(\mathbf{y}_t\right)$ and $\mathbb{E}\left[\mathbf{y}_t \mathbf{y}_t^T\right]$ that are incorrect. But note that CLS uses such a moment in its estimator, so if we blindly run CLS on $\mathbf{y}$ in place of $\mathbf{n}$, the asymptotic result will be:

$$P_{\text{CLS}} = \big(\Lambda_t + \operatorname{Var}\left(\epsilon_t\right)\big)^{-1} \Lambda_{t,t+1} \neq P.$$

Thus, CLS is not consistent in this noise model. In contrast, the simpler model we propose here uses only the mean vector and the time-lagged second moments, so it is consistent without modification when $\mathbf{n}$ is replaced by $\mathbf{y}$.

General Noise Models. The following proposition delineates a much broader class of noise models for which we can recover the needed moments of $\mathbf{n}$ from those of $\mathbf{y}$.

Proposition 3. *Suppose the noise model $p(\mathbf{y} \mid \mathbf{n})$ satisfies the following two conditions:*

(i) $\mathbf{y}_t \perp\!\!\!\perp \mathbf{y}_s \mid \mathbf{n}_t$ *for $s \neq t$ (independent noise),*

(ii) $\mathbb{E}\left[\mathbf{y}_t \mid \mathbf{n}_t\right] = A_t \mathbf{n}_t$, *for known, invertible matrix A_t.*

Then the following moments of $\mathbf{n}$ can be recovered from those of $\mathbf{y}$:

(i) $\mathbb{E}\left[\mathbf{n}_t\right] = A_t^{-1} \mathbb{E}\left[\mathbf{y}_t\right]$ *for all t,*

(ii) $\operatorname{Cov}(\mathbf{n}_s, \mathbf{n}_t) = A_s^{-1} \operatorname{Cov}(\mathbf{y}_s, \mathbf{y}_t) A_t^{-T}$ *for $s \neq t$,*

(iii) $\mathbb{E}\left[\mathbf{n}_s \mathbf{n}_t^T\right] = A_s^{-1} \mathbb{E}\left[\mathbf{y}_s \mathbf{y}_t^T\right] A_t^{-T}$ *for $s \neq t$.*

The proof can be found in the supplementary material. Note that the formulas for recovering the moments use very little information about the noise model—only the linear form of the conditional mean $\mathbb{E}\left[\mathbf{y}_t \mid \mathbf{n}_t\right]$—so very detailed knowledge of the noise mechanism is not necessary. Also note that Proposition 3 does *not* give formulas for $\operatorname{Var}\left(\mathbf{n}_t\right)$ and $\mathbb{E}\left[\mathbf{n}_t\mathbf{n}_t^T\right]$; as in our simple preceding example, these are more complicated, and require greater knowledge of the noise model than the conditional mean. Proposition 3 can be further generalized to the case where there is an affine relationship $\mathbb{E}\left[\mathbf{y}_t \mid \mathbf{n}_t\right] = A_t\mathbf{n}_t + \mathbf{b}_t$, but the expressions become more complicated.

This proposition suggests that the matrix A_t must be known in advance. For many noise models, however, it is possible to consistently estimate this matrix from the data: our experiments demonstrate this for both binomial and additive noise models.

4.1 Examples

We now give several examples of noise models that meet the conditions of Proposition 3 and lead to simple ways to recover the moments of $\mathbf{n}$ from those of $\mathbf{y}$.

Binomial or Poisson noise. A simple example is the one in which each individual is observed independently with probability α. A small variant is the case when the number of individuals counted in each state is a Poisson random variable with mean proportional to the true number. These noise models are given by:

$$y_t(i) \mid n_t(i) \sim \operatorname{Binomial}\big(n_t(i), \alpha\big)$$
$$y_t(i) \mid n_t(i) \sim \operatorname{Poisson}\big(\alpha \cdot n_t(i)\big).$$

In both cases, we have $\mathbb{E}\left[\mathbf{y}_t \mid \mathbf{n}_t\right] = \alpha\mathbf{n}_t$, which satisfies the conditions of Proposition 3 with $A_t = \alpha I$ for all t. Thus, we can recover the moments of $\mathbf{n}$ as:

$$\mathbb{E}\left[\mathbf{n}_t\right] = \alpha^{-1}\mathbb{E}\left[\mathbf{y}_t\right]$$
$$\mathbb{E}\left[\mathbf{n}_t\mathbf{n}_{t+1}^T\right] = \alpha^{-2}\mathbb{E}\left[\mathbf{y}_t\mathbf{y}_{t+1}^T\right]$$
$$\operatorname{Cov}(\mathbf{n}_t, \mathbf{n}_{t+1}) = \alpha^{-2}\operatorname{Cov}(\mathbf{y}_t, \mathbf{y}_{t+1}).$$

Additive Noise. We have already discussed additive noise models of the form $\mathbf{y}_t = \mathbf{n}_t + \epsilon_t$, for example:

$$\epsilon_t(i) \sim \operatorname{Normal}(0, \sigma^2), \quad \epsilon_t(i) \sim \operatorname{Laplace}(b).$$

These are special cases of Proposition 3 with $A_t = I$, so we can just substitute $\mathbf{y}$ for $\mathbf{n}$ in our estimator. Gaussian noise is a very common model for measurement error. Both Gaussian and Laplace noise are used in mechanisms that explicitly add noise to count data such as ours prior to release to guarantee differential privacy (Dwork and Roth, 2013; Sun et al., 2015). Our

---

Algorithm 1: Method of Moments with Noise

Input : noise model matrix A, population size N, data vectors $\mathbf{y}_t^{(k)}$ for $t = 1, \dots, T$ and $k = 1, \dots, K$.

Output: estimated transition matrix $\hat{P}$

1. Estimate mean of noisy data:
$$\hat{\mathbf{m}}_\mathbf{y} := \frac{1}{TK}\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbf{y}_t^{(k)}.$$

2. Estimate mean of true counts: $\hat{\mathbf{m}}_\mathbf{n} := A^{-1}\hat{\mathbf{m}}_\mathbf{y}$.

3. Normalize: $\hat{\boldsymbol{\mu}} := \hat{\mathbf{m}}_\mathbf{n}/\mathbf{1}^T\hat{\mathbf{m}}_\mathbf{n}$.

4. Estimate time-lagged covariance of true counts:
$$\hat{\Sigma} := A^{-1}\left(\frac{1}{(T-1)K}\sum_{t=1}^{T-1}\sum_{k=1}^{K}\mathbf{r}_t^{(k)}(\mathbf{r}_{t+1}^{(k)})^T\right)A^{-T}$$
where $\mathbf{r}_t^{(k)} := \mathbf{y}_t^{(k)} - \hat{\mathbf{m}}_\mathbf{y}$.

5. Estimate transition matrix:
$$\hat{P} := \operatorname{diag}(\hat{\boldsymbol{\mu}})^{-1}\big(N^{-1}\hat{\Sigma} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T\big).$$

---

results show that it is possible to consistently learn with private data under these noise mechanisms.

State-Dependent Detection Probability. Another interesting model occurs when the probability that an individual is counted varies by state. Suppose that the detection probability in state i is α_i, so that

$$y_t(i) \mid n_t(i) \sim \operatorname{Binomial}\big(n_t(i), \alpha_i\big).$$

Then we have $\mathbb{E}\left[\mathbf{y}_t \mid \mathbf{n}_t\right] = A\mathbf{n}_T$ with $A = \operatorname{diag}([\alpha_1, \dots, \alpha_S]^T)$, so we can also apply method of moments in this case (as long as the detection probabilities are known or can be estimated).

4.2 Putting It Together: Estimation with Noisy Data

The detailed procedure for method of moments with noisy data is given in Algorithm 1. In this algorithm, we assume that the entire process is stationary, including the noise model, so there is a single matrix A such that $\mathbb{E}\left[\mathbf{y}_t \mid \mathbf{n}_t\right] = A\mathbf{n}_t$ for all t. The algorithm accepts N and A as inputs. With exact observations, these parameters are known: $A = I$ and N is the total count at any time step. With noisy observations, A and N are not known, but can often be estimated easily. For example, in the case of binomial noise, let $z_t = \sum_{i=1}^{S}y_t(i)$ be the total number of individuals observed at time t. Then the z_t variables are iid $\operatorname{Binomial}(N, \alpha)$ random variables, from which N and α can be consistently estimated by various methods (Blumenthal and Dahiya, 1981).

4.3 Non-Stationary Processes

Our ideas can also be extended to non-stationary Markov chains, e.g., when the individual chains are started from arbitrary distributions, or when the transition probabilities are time-varying. To see this, note that Equation (6) shows that we can recover the (potentially time-varying) transition matrix P_t from the moments $\boldsymbol{\mu}_t$ and $\Sigma_{t,t+1}$ at time t for *any* t. If these moments are time-varying, instead of averaging over all t as we do in Algorithm 1, we can construct separate estimates $\hat{P}_t$ at each time step and then combine those to obtain a final parameter estimate. This makes the most sense when parameters defining the transition probabilities are shared across time steps. One such example is the case when P is time-homogenous but the chains are started in non-stationary distributions. In this case, we can estimate P by averaging the time-specific estimates: $\hat{P} = \frac{1}{T}\sum_{t=1}^{T}\hat{P}_t$. Another example is when transition probabilities are time-varying but compactly parameterized as $P_t(\boldsymbol{\theta})$ for a finite-dimensional $\boldsymbol{\theta}$. In this case, we can solve for the best parameters in a least squares sense: $\hat{\boldsymbol{\theta}} = \mathrm{argmin}_{\boldsymbol{\theta}}\sum_{t=1}^{T}\|P_t(\boldsymbol{\theta}) - \hat{P}_t\|^2$.

5 Theoretical Analysis

We now analyze the method of moments estimator to provide theoretical guarantees on its performance. We prove consistency in two different settings: the *time average* setting, when $T \to \infty$ for fixed K, and the *ensemble average* setting, when $K \to \infty$ for fixed T. We also show that the mean squared estimation error for the moments decays as $\mathcal{O}(1/TK)$.

Consistency   Fix any value of N and let $\hat{P}_{T,K}$ be the estimator of Algorithm 1 when the process is observed for T time steps and K independent realizations. Assume the noise model $p(\mathbf{y} \mid \mathbf{n})$ satisfies the conditions of Proposition 3 and that the noise model matrix A and the population size are known. (If they are not known, but can be consistently estimated, which is usually the case, then the following results still hold.) We have the following result.

Theorem 1. *The estimator of Algorithm 1 is consistent ($\hat{P}_{T,K}$ converges in probability to P) as one or both of T and K go to infinity.*

Proof. When $K \to \infty$, this is a simple consequence of the law of large numbers. We can view the sample moments in Algorithm 1 as averaging first over k and then t. For example:

$$\hat{\mathbf{m}}_{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{m}}_{\mathbf{y},t}, \quad \hat{\mathbf{m}}_{\mathbf{y},t} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{y}_t^{(k)}.$$

Since $\hat{\mathbf{m}}_{\mathbf{y},t}$ is a sample average over the K independent realizations, it converges to $\mathbf{m}_t = \mathbb{E}[\mathbf{y}_t]$. Furthermore, since the process is stationary, all terms $\hat{\mathbf{m}}_{\mathbf{y},t}$ in the time average converge to the common value $\mathbf{m} = \mathbb{E}[\mathbf{y}_1]$. Thus $\hat{\mathbf{m}}_{\mathbf{y}}$ also converges to $\mathbf{m}$. The proof that the sample covariance $\frac{1}{(T-1)K}\sum_{t,k}\mathbf{r}_t^{(k)}(\mathbf{r}_t^{(k)})^T$ converges to $\mathrm{Cov}(\mathbf{y}_1, \mathbf{y}_2)$ is similar. The estimate $\hat{P}_{T,K}$ is a deterministic function of these two sample moments, and so their convergence guarantees that $\hat{P}_{T,K}$ converges to P as $K \to \infty$.

To prove consistency when $T \to \infty$, it is clearly enough to consider the case $K = 1$, which we do now. We must argue that the time averages $\hat{\mathbf{m}}_{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t$ and $\hat{\Sigma}_{\mathbf{y}} = \frac{1}{T-1}\sum_{t=1}^{T}(\mathbf{y}_t - \hat{\mathbf{m}}_{\mathbf{y}})(\mathbf{y}_t - \hat{\mathbf{m}}_{\mathbf{y}})^T$ of the stationary process $\{\mathbf{y}_t\}$ converge to the true moments $\mathbb{E}[\mathbf{y}_1]$ and $\mathrm{Cov}(\mathbf{y}_1, \mathbf{y}_2)$ as $T \to \infty$, which requires reasoning about ergodic properties of the process. A process is called *mean-ergodic* if its time average converges to the population mean (Papoulis, 1991): applied to our problem, this is exactly equivalent to $\hat{\mathbf{m}}_{\mathbf{y}}$ being a consistent estimator of $\mathbb{E}[\mathbf{y}_1]$. Similarly, the covariance estimate $\hat{\Sigma}$ will converge to the population covariance if and only if the process $\{\mathbf{Z}_t\}$ is mean-ergodic, where $\mathbf{Z}_t = \mathbf{y}_t\mathbf{y}_{t+1}^T$. We will focus on proving that $\{\mathbf{Z}_t\}$ is mean-ergodic, which is a less obvious property than that of $\{\mathbf{y}_t\}$. It is enough to show that $\{z_t\}$ is mean-ergodic for an arbitrarily chosen entry of $z_t = y_t(i)y_{t+1}(j)$ of $\mathbf{Z}_t$. Furthermore, we will simplify the issue slightly by showing instead that this is true for $z_t = n_t(i)n_{t+1}(j)$. This is justified by our noise model: since noise is independent at each time step (condition (i) of Proposition 3), one can show that the ergodic properties of the $\{\mathbf{y}_t\}$ process follow from those of the $\{\mathbf{n}_t\}$ process.

Let $\gamma(k) = \mathrm{Cov}(z_t, z_{t+k})$ be the autocovariance function of the process $\{z_t\}$. A sufficient condition for $\{z_t\}$ to be mean-ergodic is (Papoulis, 1991):

$$\lim_{k \to \infty} \gamma(k) = 0 \qquad (7)$$

We give a detailed verification of this condition in the supplementary material. Intuitively, it is rather clear why this condition holds. Since each individual follows an ergodic Markov chain, the state x_{t+k} becomes independent of x_t as the time lag k goes to infinity, and thus the autocovariance $\mathrm{Cov}([x_t = i], [x_{t+k} = i'])$ goes to zero for any state-pair (i, i'). Equation (7) asserts that an analogous autocovariance decay holds for products $n_t(i)n_{t+1}(j)$ of population counts at two adjacent time steps. $\square$

Convergence Rates   The consistency arguments in the previous section can also be modified to obtain convergence rates.
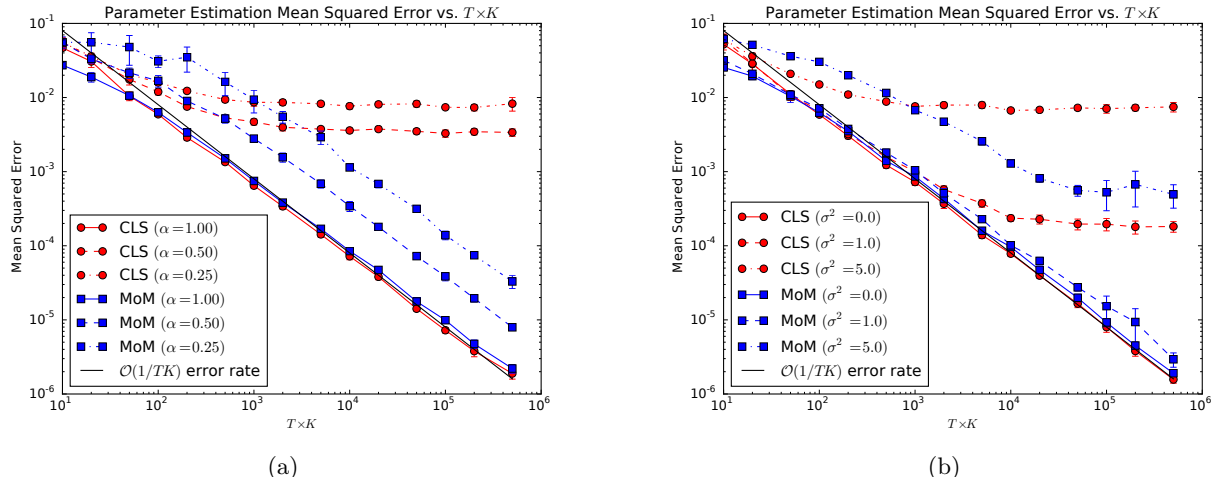
Figure 2: Parameter estimation mean squared error of MoM and CLS vs. $T \times K$ for (a) binomial noise and (b) additive Gaussian noise.

Theorem 2. *As the product $TK \to \infty$, the estimates $\hat{\mathbf{m}}_{\mathbf{y}}$ and $\hat{\Sigma}$ are unbiased and have variance (and thus mean squared error) $\mathcal{O}(1/TK)$.*

Proof. We can modify our consistency argument above to compute the variance σ^2 of the time average estimate for $n_t(i)n_t(j)$ as $T \to \infty$ for a single chain ($K = 1$), according to the following formula (Papoulis, 1991):

$$\sigma^2 = \frac{1}{T} \sum_{k=-(T-1)}^{T-1} \left(1 - \frac{|k|}{T}\right)\gamma(k)$$

We show in the supplementary material that $|\gamma(k)| \leq C\alpha^k$ for some constants $\alpha \in (0, 1)$ and $C > 0$, which implies that the sum above remains finite as $T \to \infty$ and thus σ^2 is $\mathcal{O}(1/T)$. Now, the final estimate is an average over K iid estimators (one for each chain) with variance σ^2. Therefore, it has variance $\sigma^2/K = \mathcal{O}(1/TK)$. $\qquad\square$

We have not presented a theoretical analysis of how the error rates of the moment estimates combine to give an error rate for the final parameter estimate $\hat{P}$. Our experiments present evidence that the the mean squared error of $\hat{P}$ is $\mathcal{O}(1/TK)$.

6 Experiments

We conduct a number of experiments to validate our theoretical results and examine convergence rates of CLS, method of moments (hereafter, MoM), and additional baselines.

Setup. We generate random $S \times S$ transition matrices P by drawing each row from the Dirichlet distribution using the mean and precision parameterization (Minka, 2000) with mean vector $\mathbf{1}/S$ and precision D. In preliminary experiments, we varied D and S to examine the effectiveness of estimators for a broad range of qualitatively different transition models. For low D, the transitions from any given state are concentrated on a few states, so there is high dependence between time steps. For high D, each row of P is very close to the mean, which implies that time steps are nearly independent. These experiments revealed that trends in estimation error of MoM and CLS are very consistent over a broad range of S and D; therefore, we report results here only for $S = 10$ and $D = 0.5$.

Once the transition model P is fixed, we compute the stationary distribution $\boldsymbol{\pi}$ and generate the aggregate data by drawing $\mathbf{n}_1 \sim \text{Multinomial}(N, \boldsymbol{\pi})$ and then simulating the aggregate Markov process to generate $\mathbf{n}_2, \dots, \mathbf{n}_T$. We then generate the noisy vectors $\mathbf{y}_1, \dots, \mathbf{y}_T$ using the binomial and Gaussian noise models (parameters α and σ^2) described in Section 4.1. We repeat the entire process K times to generate the final observed data $\{\mathbf{y}_t^{(k)}\}$.

Baselines and Evaluation. Our main experiments compare MoM to CLS. We also compare to the *limited information maximum likelihood estimator* (LIMLE) of (MacRae, 1977), which was explicitly designed for noisy observations. LIMLE, however, assumes the marginals of process are time-varying and provide enough information to estimate the parameters: it finds P to maximize the likelihood of the approximate model where $\mathbf{n}_t \sim \text{Multinomial}(N, \boldsymbol{\mu}_t)$ independently for all t. We evaluate the quality of each estimator $\hat{P}$ using entrywise mean squared error: $\frac{1}{S^2}\|\hat{P} - P\|_F^2$.
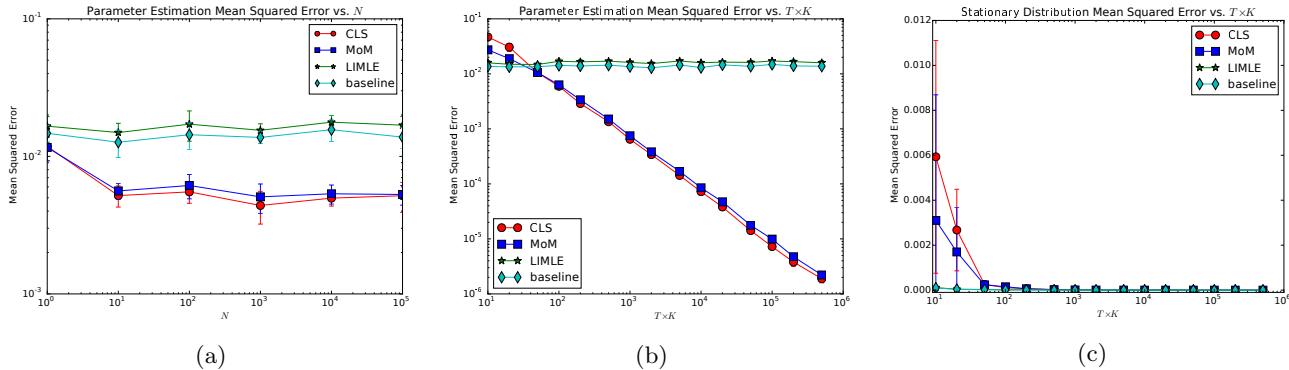
Figure 3: (a) Effect of population size N on parameter estimation error (perfect observations, $K = 1, T = 100$). (b, c) Mean squared error vs. $T \times K$ for (b) parameter estimation and (c) stationary distribution estimation.

Consistency and Convergence Rates. We first examined the asymptotic behavior of MoM and CLS as T and K grow by running both estimators for all combinations of $T \in \{10, 100, 1000, 10000\}$ and $K \in \{1, 2, 5, 10, 20, 50\}$. We repeated each combination 10 times. Figures 2a and 2b show estimation error for MoM and CLS plotted against the product $T \times K$ for binomial and Gaussian noise, respectively. Each data point averages over all trials for all combinations of T and K that yield the same product (error bars indicate the 95% confidence interval of the mean error over that set of trials). For this experiment, we fixed $N = 100$; we will later see that error has almost no dependence on N (Figure 3a).

In Figures 2a and 2b, we see that the error of both MoM and CLS with perfect observations ($\alpha = 1.00, \sigma^2 = 0$) decays almost exactly as $1/TK$. We also see very clearly that CLS is not consistent with noisy observations; the estimation error flattens out at TK grows, and at higher error levels when there is more noise. For binomial noise and Gaussian noise with $\sigma^2 = 1.0$, MoM retains its $1/TK$ convergence rate. For Gaussian noise with $\sigma^2 = 5.0$, the convergence of MoM appears to slow down for large $T \times K$; it may be the case, when the noise level is high enough, that the convergence rate of the final parameter estimate $\hat{P}$ is different from the $1/TK$ rate that we proved for the moment estimates.

Effect of N. Figure 3a shows the estimation error vs. the population size N for each method. Perhaps surprisingly, N has very little effect on estimation quality. Therefore, if we only observe aggregate data, there is no loss in estimation accuracy as the population size grows (a population of size $N = 1$ is the same as one of size $N = 10^6$).

Individual vs. Aggregate Data. We can now address one of our questions from the outset: How does estimation with aggregate data compare to estimation with individual data? Suppose the total number of individuals is fixed and equal to M, and we consider the alternatives of aggregating the data into one population of size M (i.e. $K = 1, N = M$) vs. observing individual data ($K = M, N = 1$). Our results show that error decays as $O(1/TK)$ independently of N, so the error in the former case is approximately M times the latter. In other words, estimation error increases by a factor of M when we move from individual data to aggregate data in a population of size M; we would therefore need to observe the aggregate data of a system for a factor of M additional time steps or M additional independent realizations to achieve estimation quality comparable to observing individual data.

Comparison to LIMLE. It is fairly clear that LIMLE, which approximates the model by one where each time step is independent, will fail to recover P when the process is stationary. This is because it cannot distinguish between two transition models that yield the same stationary distribution, and hence the same marginals at each time step. Figures 3b confirms very clearly that LIMLE is not consistent. In fact, we can characterize LIMLE as searching for *any* transition matrix P that has the correct stationary distribution $\boldsymbol{\pi}$. To verify this, we do two things. First, we introduce a naive baseline that does exactly the same thing: it estimates the stationary distribution $\hat{\boldsymbol{\pi}}$ by averaging over t and k, and then sets $\hat{P}$ to the matrix with each row equal to $\hat{\boldsymbol{\pi}}$. (The result is a model where x_t is an independent draw from $\hat{\boldsymbol{\pi}}$ at each time step). Figure 3b shows that this baseline actually outperforms LIMLE slightly in our stationary setting. Second, we compare all methods in terms of their error estimating the correct *stationary distribution*. Figure 3c shows that LIMLE indeed quickly converges to the correct stationary distribution, even though it fails to estimate P. CLS and MoM are slightly slower to converge to the correct stationary distribution, but consistently estimate P.

References

Dennis J. Aigner and Stephen M. Goldfeld. Estimation and prediction from aggregate data when aggregates are measured more accurately than their components. *Econometrica: Journal of the Econometric Society*, 1974.

Saul Blumenthal and Ram C. Dahiya. Estimating the binomial parameter n. *Journal of the American Statistical Association*, 1981.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2013.

Seth R. Flaxman, Yu-Xiang Wang, and Alexander J. Smola. Who supported Obama in 2012?: Ecological inference through distribution regression. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

Matthew T. Jones. Estimating Markov transition matrices using proportions data: an application to credit risk. *International Monetary Fund (IMF) Working Paper*, 2005.

John D. Kalbfleisch, Jerald F. Lawless, and William M. Vollmer. Estimation in Markov models from aggregate data. *Biometrics*, 1983.

Gary King. *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data.* Princeton University Press, 2013.

Tsoung-Chao Lee, George G. Judge, and Arnold Zellner. *Estimating the parameters of the Markov probability model from aggregate time series data.* North-Holland Pub. Co., 1970.

David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov chains and mixing times.* American Mathematical Soc., 2009.

Li-Ping Liu, Daniel R. Sheldon, and Thomas G. Dietterich. Gaussian approximation of collective graphical models. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

Elizabeth C. MacRae. Estimation of time-varying markov processes with aggregate data. *Econometrica: journal of the Econometric Society*, 1977.

Albert Madansky. Least squares estimation in finite Markov processes. *Psychometrika*, 1959.

George A. Miller. Finite markov processes in psychology. *Psychometrika*, 17(2):149–167, 1952.

Thomas P. Minka. Estimating a Dirichlet distribution. Technical report, MIT, 2000.

Athanasios Papoulis. *Probability, random variables, and stochastic processes.* McGraw-Hill, Inc., 1991.

Daniel R. Sheldon and Thomas G. Dietterich. Collective graphical models. *Neural Information Processing Systems (NIPS)*, 2011.

Daniel R. Sheldon, Tao Sun, Akshat Kumar, and Thomas G. Dietterich. Approximate inference in collective graphical models. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann. Math. Statist.*, 21(1):124–127, 03 1950.

Tao Sun, Daniel R. Sheldon, and Akshat Kumar. Message passing for collective graphical models. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Adriaan P. Van Der Plas. On the estimation of the parameters of Markov probability models using macro data. *The Annals of Statistics*, 1983.

A   Extra Proofs

A.1   Additional Details for Proof of Proposition 2

We wish to show that $\Lambda_t = N\big(\text{diag}(\boldsymbol{\mu}_t) + (N-1)\boldsymbol{\mu}_t\boldsymbol{\mu}_t^T\big)$ is invertible. The Sherman-Morrison formula (Sherman and Morrison, 1950) gives the inverse of a matrix that is equal to an invertible matrix ($\text{diag}(\boldsymbol{\mu}_t)$) plus a rank-one matrix (the rank-one outer product of $\boldsymbol{\mu}_t$).

Specifically, let $D = \text{diag}(\boldsymbol{\mu}_t)$ and then we have

$$
\begin{aligned}
\Lambda_t^{-1} &= N^{-1}\big(D + (N-1)\boldsymbol{\mu}_t\boldsymbol{\mu}_t^T\big)^{-1} \\
&= N^{-1}\Big(D^{-1} - \frac{(N-1)D^{-1}\boldsymbol{\mu}_t\boldsymbol{\mu}_t^T D^{-1}}{1 + (N-1)\boldsymbol{\mu}^T D^{-1}\boldsymbol{\mu}_t}\Big) \\
&= N^{-1}\Big(D^{-1} - \frac{N-1}{N}\mathbf{1}\mathbf{1}^T\Big).
\end{aligned}
$$

We have used the fact that $D^{-1}\boldsymbol{\mu}_t = \mathbf{1}$ (the all ones vector) and that $\mathbf{1}^T\boldsymbol{\mu}_t = 1$, since $\boldsymbol{\mu}_t$ is a vector of marginal probabilities.

A.2   Proof of Proposition 3

Proof. Using condition (ii) of the proposition, we write:

$$
\mathbb{E}\left[\mathbf{y}_t\right] = \mathbb{E}\big[\mathbb{E}\left[\mathbf{y}_t|\mathbf{n}_t\right]\big] = \mathbb{E}\left[A_t\mathbf{n}_t\right] = A_t\mathbb{E}\left[\mathbf{n}_t\right].
$$

Since A_t is invertible, we have $\mathbb{E}\left[\mathbf{n}_t\right] = A_t^{-1}\mathbb{E}\left[\mathbf{y}_t\right]$, which proves conclusion (i).

For the non-central second moments, we have for $s \neq t$:

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{y}_s\mathbf{y}_t^T\right] &= \mathbb{E}\big[\mathbb{E}\left[\mathbf{y}_s\mathbf{y}_t^T|\mathbf{n}_s, \mathbf{n}_t\right]\big] \\
&= \mathbb{E}\big[\mathbb{E}\left[\mathbf{y}_s|\mathbf{n}_s\right] \cdot \mathbb{E}\left[\mathbf{y}_t^T|\mathbf{n}_t\right]\big] \\
&= \mathbb{E}\left[(A_s\mathbf{n}_s)(A_t\mathbf{n}_t)^T\right] \\
&= A_s\mathbb{E}\left[\mathbf{n}_s\mathbf{n}_t^T\right]A_t^T
\end{aligned}
\tag{8}
$$

The second line uses condition (i) of the proposition: $\mathbf{y}_s$ and $\mathbf{y}_t$ are conditionally independent given $\mathbf{n}_s$ and $\mathbf{n}_t$ if $s \neq t$. Therefore, we have $\mathbb{E}\left[\mathbf{n}_s\mathbf{n}_t^T\right] = A_s^{-1}\mathbb{E}\left[\mathbf{y}_s\mathbf{y}_t^T\right]A_t^{-T}$, which proves conclusion (iii).

For conclusion (ii), we have for $s \neq t$:

$$
\begin{aligned}
\text{Cov}(\mathbf{y}_s, \mathbf{y}_t) &= \mathbb{E}\left[\mathbf{y}_s\mathbf{y}_t^T\right] - \mathbb{E}\left[\mathbf{y}_s\right]\mathbb{E}\left[\mathbf{y}_t\right]^T \\
&= A_s\mathbb{E}\left[\mathbf{n}_s\mathbf{n}_t^T\right]A_t^T - A_s\mathbb{E}\left[\mathbf{n}_s\right]\mathbb{E}\left[\mathbf{n}_t\right]^T A_t^T \\
&= A_s\big(\mathbb{E}\left[\mathbf{n}_s\mathbf{n}_t^T\right] - \mathbb{E}\left[\mathbf{n}_s\right]\mathbb{E}\left[\mathbf{n}_t\right]^T\big)A_t^T \\
&= A_s\text{Cov}(\mathbf{n}_s, \mathbf{n}_t)A_t^T,
\end{aligned}
$$

so that $\text{Cov}(\mathbf{n}_s, \mathbf{n}_t) = A_s^{-1}\text{Cov}(\mathbf{y}_s, \mathbf{y}_t)A_t^{-T}$, which completes the proof. $\square$

A.3   Additional Details for Proof of Theorem 1

We wish to show that $\lim_{k\to\infty}\gamma(k) = 0$, where $\gamma(k) = \text{Cov}\big(n_t(i)n_{t+1}(j), n_{t+k}(i)n_{t+k+1}(j)\big)$. We have

$$\gamma(k) = \mathrm{Cov}(n_t(i)n_{t+1}(j), n_{t+k}(i)n_{t+k+1}(j))$$

$$= \mathrm{Cov}\left(\sum_{a=1}^{N}\sum_{b=1}^{N}[x_t^{(a)} = i][x_{t+1}^{(b)} = j], \ \sum_{c=1}^{N}\sum_{d=1}^{N}[x_{t+k}^{(c)} = i][x_{t+k+1}^{(d)} = j] \right)$$

$$= \sum_{a,b,c,d=1}^{N} \mathrm{Cov}\left([x_t^{(a)} = i][x_{t+1}^{(b)} = j], \ [x_{t+k}^{(c)} = i][x_{t+k+1}^{(d)} = j] \right)$$

It is enough to show that the covariance in the summand goes to zero for any choice of four individuals $a, b, c, d \in \{1, \ldots, N\}$. Clearly, it is equal to zero when the individuals $\{a, b\}$ do not overlap with $\{c, d\}$, because individuals are independent. We will verify that the covariance goes to zero for the choice $a = b = c = d := m$, which, since it involves only a single individual, is the case with the *greatest* dependence between times t and $t+k$. Verifying the statement for other combinations of a, b, c, d is similar. Because we are considering a single individual m, we now drop the superscript and write $x_t := x_t^{(m)}$. We can rewrite the covariance as:

$$\mathrm{Cov}\left([x_t = i][x_{t+1} = j], \ [x_{t+k} = i][x_{t+k+1} = j] \right)$$

$$= \mathbb{E}\left[[x_t = i][x_{t+1} = j][x_{t+k} = i][x_{t+k+1} = j] \right] - \mathbb{E}\left[[x_t = i][x_{t+1} = j] \right]\mathbb{E}\left[[x_{t+k} = i][x_{t+k+1} = j] \right]$$

$$= \mathrm{Pr}\left(x_t = i, x_{t+1} = j, x_{t+k} = i, x_{t+k+1} = j \right) - \mathrm{Pr}\left(x_t = i, x_{t+1} = j \right)\mathrm{Pr}\left(x_{t+k} = i, x_{t+k+1} = j \right)$$

$$= \mu(i,j) \cdot (P^{k-1})_{ji} \cdot P(i,j) \ - \ \mu(i,j)^2 . \tag{9}$$

In the last line, we apply several facts about the Markov chain. Here, $\mu(i,j) = \mathrm{Pr}\left(x_t = i, x_{t+1} = j \right)$ is the (time-independent) pairwise marginal, $(P^{k-1})_{ji} = \mathrm{Pr}\left(x_{t+k} = i \mid x_{t+1} = j \right)$ and $P(i,j) = \mathrm{Pr}\left(x_{t+k+1} = j \mid x_{t+k} = i \right)$. Since the Markov chain is ergodic, $\lim_{k \to \infty}(P^{k-1})_{ji} = \pi(i)$, so the first term of Equation (9) becomes:

$$\lim_{k \to \infty} \mu(i,j)(P^{k-1})_{ji}P(i,j) = \mu(i,j)\pi(i)P(i,j) = \mu(i,j)^2.$$

Putting it all together, we see that the limit as k goes to infinity of the covariance in Equation (9) is $\mu(i,j)^2 - \mu(i,j)^2 = 0$, as desired. This completes the proof.

A.4 Additional Details for Proof of Theorem 2

We wish to show that $|\gamma(k)|$ decays exponentially to zero as $k \to \infty$, where $|\gamma(k)| = |\mathrm{Cov}\left(n_t(i)n_{t+1}(j), n_{t+k}(i)n_{t+k+1}(j) \right)|$. We follow the exact same steps in Section A.3 up through Equation (9) where we instead desire $|(P^{k-1})_{ji} - \pi(i)| \leq C\alpha^k$ for some constants $\alpha \in (0,1)$ and $C > 0$. This is proved for irreducible and aperiodic P as Theorem 4.9 in (Levin et al., 2009). Using this fact together with Equation (9), we have:

$$\left| \mathrm{Cov}\left([x_t = i][x_{t+1} = j], \ [x_{t+k} = i][x_{t+k+1} = j] \right) \right| = \left| \mu(i,j) \cdot (P^{k-1})_{ji} \cdot P(i,j) \ - \ \mu(i,j)^2 \right|$$

$$= \left| \mu(i,j) \cdot (P^{k-1})_{ji} \cdot P(i,j) \ - \ \mu(i,j) \cdot \pi(i) \cdot P(i,j) \right|$$

$$= \mu(i,j) \cdot \left| (P^{k-1})_{ji} - \pi(i) \right| \cdot P(i,j)$$

$$\leq \mu(i,j) \cdot C\alpha^k \cdot P(i,j)$$

$$= C'\alpha^k.$$

for $C' = \mu(i,j)P(i,j)C$. Thus the result is proved.