Maximum a Posteriori Estimation for Markov Chains Based on Gaussian Markov Random Fields

Hao Wu and Frank Noé

Abstract

In this paper, we present a Gaussian Markov random field (GMRF) model for the transition matrices (TMs) of Markov chains (MCs) by assuming the existence of a neighborhood relationship between states, and develop the maximum a posteriori (MAP) estimators under different observation conditions. Unlike earlier work on TM estimation, our method can make full use of the similarity between different states to improve the estimated accuracy, and the estimator can be performed very efficiently by solving a convex programming problem. In addition, we discuss the parameter choice of the proposed model, and introduce a Monte Carlo cross validation (MCCV) method. The numerical simulations of a diffusion process are employed to show the effectiveness of the proposed models and algorithms.

1 Introduction

Markov chain (MC) models provide a general modeling framework for describing state evolutions of stochastic and memoryless systems, and are now important and powerful tools for an enormous range of mathematical applications, including science, economics, and engineering. Here we only focus on the finite discrete-time homogeneous MC model, which is one of the most common MC models, and whose dynamics can be simply characterized by a transition matrix (TM) $\boldsymbol{T} = [T_{ij}] \in \mathbb{R}^{n\times n}$ with T_{ij} the transition probability from the i-th state to the j-th state. Such models arise in molecular systems [1], economic forecasts and evaluations [2, 3], web navigation [4], quantitative analysis of sports [5] and many others. In most applications, the main problem is to estimate the transition probabilities from observed data.

In the past few decades, a lot of different techniques have been proposed to estimated the TMs. Many early researches devoted to the least-square (LS) approaches [6, 7, 8], for MC models can be transformed to linear stochastic systems with zero-mean noise. The advantages of these methods are that there are well developed theories and algorithms for LS regression, and they can easily extended to the case where only sample proportions are available from time series data. However, the conventional LS estimators may violate the nonnegative constraints on TMs. Thus, some restricted LS methods [9, 10, 11] based on constrained quadratic programming algorithms were developed to avoid this problem. Some researchers [7, 10] suggested utilizing the weighted LS and weighted restricted methods to solve the problem of heteroscedasticity.

By now, the best known and most popular estimation method of MC models is maximum likelihood (ML) estimator which was proposed in [12], for it is consistent and asymptotically normally distributed as the sample size increases [13], and can be efficiently calculated by counting transition pairs. Some experiments show ML estimator is superior to the LS estimators [14]. Moreover, the ML method can be applied to estimation from aggregate time series data [15], and revsersible TM estimation for some physical and chemical processes [1].

Recently, the Bayesian approach [16, 15] to TM estimation has received a good deal of attention. In this approach, an unknown TM is assumed to be a realization of some prior model, and the posterior distribution given observed data can be obtained by Bayes' rule. Comparing to the non-Bayesian methods, the Bayesian estimator can provide much more information, such as confidence intervals and expectations of some functions of the TM, than a single point estimate, and is more reliable for small size data set if the prior model is appropriately designed. The most commonly used prior distribution is the matrix beta distribution with density

$$p\left(\boldsymbol{T}|\boldsymbol{\Theta}\right) \propto \prod_{i,j} T_{ij}^{\theta_{ij}-1}$$

where $\boldsymbol{\Theta} = [\theta_{ij}]$ is a nonnegative parameter matrix. It is a conjugate prior for the likelihood of $\boldsymbol{T}$ and can be easily analyzed and efficiently sampled since each row of $\boldsymbol{T}$ follows the Dirichlet distribution. In

some applications, $\boldsymbol{\Theta} = \mathbf{1}$ and $\boldsymbol{\Theta} = \mathbf{0}$ are recommended, because $p(\boldsymbol{T}|\boldsymbol{\Theta})$ is equivalent to the uniform distribution when $\boldsymbol{\Theta} = \mathbf{1}$ [17], and $\boldsymbol{\Theta} = \mathbf{0}$ makes the posterior mean of the TM identical to the ML estimate [18]. The matrix $\boldsymbol{\Theta}$ can also be optimized by using the empirical Bayes approach [19]. The matrix beta prior distribution based Bayesian estimation of revsersible TM was investigated in [17]. The shortcoming of the matrix beta prior is that it does not take into account possible correlations between different rows of the transition matrix. Assoudou and Essebbar [20, 21] proposed the Jeffreys' prior (a non-informative prior) model for TMs to overcome this problem, and no extra parameter is required in this model. However, the Jeffreys' prior distribution is too complicated for deriving the Bayesian estimator, and can only be applied to MC models with very few states in practice.

The major objective of this paper is to propose a new prior model for MCs based on the Gaussian Markov random filed (GMRF). The GMRF [22, 23, 24, 25] model is a specific Gaussian field model, and frequently used in spatial statistics and image processing, which constructs a global distribution of a spatial function by considering the local correlations between points or regions. In this paper, we assume that the state space of the MC has neighborhood structure and the adjacent states have similar transition behaviors. This assumption generally holds for the grid based approximate models of continuous space MCs, and the case that the state space has a distance metric. A GMRF prior model of TMs is then designed according to the assumption, and the corresponding maximum a posteriori (MAP) estimator is developed. In comparison with the existing models, the new prior model is able to utilize the similarity relationship of states better. Although the new prior is not a conjugate prior, the MAP estimate can be calculated efficiently by convex programming, and there is only one extra parameter is required, which can be selected by the cross validation (CV) method. Moreover the estimation problem with noisy data is considered, and the expectation maximization (EM) algorithm is used to get the MAP estimate.

2 Background

2.1 Gaussian Markov Random Fields

Let $G = (V, E)$ be an undirected graph without loop edges, where V is the set of vertices and $E \subset V \times V$ is the edge set. And vertices $u, v \in V$ are said to be adjacent iff $(u, v) \in E$, which is denoted by $u \sim v$. It is clear that $\forall u, v \in V$, $v \nsim v$ and $u \sim v \Leftrightarrow v \sim u$. A Gaussian Markov random field (GMRF) Y on G is a Gaussian stochastic function that assigns to each vertex v a real number $Y(v)$. Here we only introduce the widely used intrinsic GMRF model [23, 26], which is often specified through the following distribution

$$p_{GMRF}(\boldsymbol{y}|\sigma) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{u\sim v}\left(\frac{Y(u)-Y(v)}{d^2(u,v)}\right)^2\right) \tag{1}$$

where $\boldsymbol{y} = \{Y(v)|v \in V\}$, σ is a parameter that controls variation in Y, and $d(\cdot,\cdot)$ denotes a distance measure between vertices. And the distribution (1) has the Markovian property that

$$\begin{aligned} p_{GMRF}(Y(v)|Y(u), u \neq v) &= p_{GMRF}(Y(v)|Y(u), u \in \partial v) \\ &= \mathcal{N}\left(Y(v)|\mu(v), \sigma^2(v)\right) \end{aligned} \tag{2}$$

where $\partial v = \{u|u \sim v, u \in V\}$, $\mathcal{N}\left(\cdot|\mu, \sigma^2\right)$ denotes the probability density function of the normal distribution with mean μ and variance σ^2, and

$$\mu(v) = \left(\sum_{u\in\partial v}\frac{Y(u)}{d^2(u,v)}\right) / \left(\sum_{u\in\partial v}\frac{1}{d^2(u,v)}\right) \tag{3}$$

$$\sigma^2(v) = \sigma^2 / \left(\sum_{u\in\partial v}\frac{1}{d^2(u,v)}\right) \tag{4}$$

That is, $Y(v)$ is conditionally independent of $\{Y(u)|u \notin \partial v\}$ for given neighbors, and the conditional distribution is centered at the weighted average of $\{Y(u)|u \in \partial v\}$.

2.2 Markov Chains

We consider a time-homogeneous Mariv chain (MC) $\{x_t|t \geq 0\}$ on the finite state space $S = \{s_1, \ldots, s_n\}$. Its probability model can be described by a transition matrix (TM) $\boldsymbol{T} = [T_{ij}] \in \mathbb{R}^{n\times n}$ whose entries are given by

$$T_{ij} = p(x_{t+1} = s_j | x_t = s_i) \tag{5}$$

where

$$\sum_j T_{ij} = 1, \quad T_{ij} > 0 \tag{6}$$

Here we define $\Omega_n = \{\boldsymbol{T} | \boldsymbol{T} \in \mathbb{R}^{n\times n}$ is a stochasic matrix$\}$, which is a convex set.

And the probability distribution of the finite-length state sequence $\{x_0, x_1, \ldots, x_m\}$ given $\boldsymbol{T}$ can be expressed as

$$p\left(x_{0:m}|\boldsymbol{T}\right) = \prod_{i,j} T_{ij}^{C_{ij}} \tag{7}$$

where entries of count matrix $\boldsymbol{C} = [C_{ij}]$ are numbers of observed transition pairs with

$$C_{ij} = \left|\left\{\left(x_t, x_{t+1}\right) | x_t = s_i, x_{t+1} = s_j, 0 \le t \le m-1\right\}\right| \tag{8}$$

3 GMRF Based MC Model Estimation

3.1 GMRF Prior

Given an MC state space $S = \{s_1, \ldots, s_n\}$, the purpose of this subsection is provide a GMRF model based prior distribution for the TM $\boldsymbol{T} = [T_{ij}]$. Assuming a neighborhood structure on the state space, which is common in the case that the discrete states are obtained by decomposing a large scale state space using some discretization or aggregation method, we construct a neighborhood relation between the transition pairs as

$$\left(s_i, s_j\right) \sim \left(s_k, s_l\right) \Leftrightarrow \left(s_i, s_j\right) \in \left(\partial s_k \cup \{s_k\}\right) \times \left(\partial s_l \cup \{s_l\}\right) \setminus \{\left(s_k, s_l\right)\} \tag{9}$$

Since the new neighborhood relation (9) is also symmetric and irreflexive, we can model the unknown matrix $\boldsymbol{T}$ by GMRF model with distribution

$$p_{GMRF}\left(\boldsymbol{T}|\sigma\right) \propto \exp\left(-u\left(\boldsymbol{T}, \sigma\right)\right) \tag{10}$$

where

$$u\left(\boldsymbol{T}, \sigma\right) = \frac{1}{2\sigma^2} \sum_{\left(s_i, s_j\right) \sim \left(s_k, s_l\right)} \left(\frac{T_{ij} - T_{kl}}{d_{ijkl}^2}\right)^2 \tag{11}$$

d_{ijkl} is the distance between $\left(s_i, s_j\right)$ and $\left(s_k, s_l\right)$, and here defined as

$$d_{ijsk} = \sqrt{d^2\left(s_i, s_k\right) + d^2\left(s_j, s_l\right)} \tag{12}$$

However, the realization of distribution (10) does not satisfy (6) in the general case. Therefore we modify the prior distribution as

$$p_{GMC}\left(\boldsymbol{T}|\sigma\right) = p_{GMRF}\left(\boldsymbol{T}|\sigma, \boldsymbol{T} \in \Omega_n\right) = \begin{cases} \frac{1}{z(\sigma)} \exp\left(-u\left(\boldsymbol{T}, \sigma\right)\right), & \boldsymbol{T} \in \Omega_n \\ 0, & \boldsymbol{T} \notin \Omega_n \end{cases} \tag{13}$$

where

$$z\left(\sigma\right) = \int_{\Omega_n} \exp\left(-u\left(\boldsymbol{T}, \sigma\right)\right) \mathrm{d}\boldsymbol{T} \tag{14}$$

i.e., $\boldsymbol{T}$ follows the distribution $p_{GMRF}\left(\boldsymbol{T}|\sigma\right)$ under the constraint that $\boldsymbol{T} \in \Omega_n$.

For a stochastic realization of $p_{GMC}\left(\boldsymbol{T}|\sigma\right)$, values of T_{ij} and T_{kl} are always close if distances between s_i and s_k, s_j and s_l are small. In comparison with the conventional prior models of transition matrices, the proposed model can describe the stochastic relationships between $p\left(x_{t+1}|x_t = s_i\right)$ and $p\left(x_{t+1}|x_t = s_k\right)$ for $i \neq k$. Further, $u\left(\boldsymbol{T}, \sigma\right)$ is a positive-semidefinite quadratic form in $\boldsymbol{T}$, which has many analytical and computational advantages.

3.2 MAP Estimation

The maximum a posteriori (MAP) estimate of the TM $\boldsymbol{T}$ of an MC from observed data $\{x_0, \ldots, x_t\}$ with count matrix $\boldsymbol{C} = [C_{ij}]$ is given by

$$\hat{\boldsymbol{T}} = \arg\max_{\boldsymbol{T}} \left\{\log p\left(\boldsymbol{C}|\boldsymbol{T}\right) + \log p\left(\boldsymbol{T}\right)\right\} \tag{15}$$

Using the proposed GMRF prior model and assuming the parameter σ is known, (15) is equivalent to the following optimization problem

$$\hat{\boldsymbol{T}}(\sigma) = \arg\min_{\boldsymbol{T}\in\Omega_n} \left\{ -\sum_{i,j} C_{ij} \log T_{ij} + u(\boldsymbol{T}, \sigma) \right\} \tag{16}$$

It is a convex problem and can be solved without any spurious local minima. We discuss how to perform the optimization efficiently in Appendix A.

Remark 1 *It is clear that the MAP estimation* (16) *is equivalent to the maximum likelihood (ML) estimator when* $\sigma \to \infty$.

3.3 Choice of σ

We now consider the case that σ is unknown. Motivated by the above analysis, it seems reasonable to jointly estimate $\boldsymbol{T}$ and σ by MAP method as

$$\left(\hat{\boldsymbol{T}}, \hat{\sigma}\right) = \arg\max_{\boldsymbol{T},\sigma} \{\log p(\boldsymbol{C}|\boldsymbol{T}) + \log p(\boldsymbol{T}, \sigma)\} \tag{17}$$

After some manipulations, (17) reduces to

$$\left(\hat{\boldsymbol{T}}, \hat{\sigma}\right) = \arg\min_{\boldsymbol{T}\in\Omega_n,\sigma} \left\{ -\sum_{i,j} C_{ij} \log T_{ij} + u(\boldsymbol{T}, \sigma) - \log p(\sigma) + \log z(\sigma) \right\} \tag{18}$$

where $p(\sigma)$ denotes the prior distribution of σ. In this form, the first three terms are easily computed, but $\log z(\sigma)$ is an intractable function of σ since it requires the computation of a integral on Ω_n.

So here we use cross-validation (CV) approach to select the value of σ, and adopt the Monte Carlo cross-validation (MCCV) method proposed in [27], which is considered an effective method of selecting models, and performs more stably than the traditional v-fold CV and leave-one-out CV [28, 29, 30, 31]. The MCCV of a σ is conducted by the following steps:

Step 1. Partition the set of observed state transition pairs randomly into train and test subsets, where the test subset is a fraction β (typically 0.5) of the overall set, and the corresponding count matrices are denoted by $\boldsymbol{C}_k^{train}$ and $\boldsymbol{C}_k^{test}$.

Step 2. Calculate

$$\hat{\boldsymbol{T}}_k(\sigma) = \arg\max_{\boldsymbol{T}\in\Omega^n} \left\{\log p\left(\boldsymbol{C}_k^{train}|\boldsymbol{T}\right) - u(\boldsymbol{T}, \sigma)\right\} \tag{19}$$

and the predictive log-likelihood

$$CV_k(\sigma) = \log p\left(\boldsymbol{C}_k^{test}|\hat{\boldsymbol{T}}_k(\sigma)\right) \tag{20}$$

Step 3. Repeat the above steps for $k = 1, \ldots, K$ and select

$$\sigma^* = \arg\max_{\sigma} CV(\sigma) \tag{21}$$

with $CV(\sigma) = \sum_k CV_k(\sigma)/K$.

It can be seen from (20) that $CV_k(\sigma) \to -\infty$ if the (i,j)-th entry of $\boldsymbol{C}_k^{test}$ is positive and that of $\hat{\boldsymbol{T}}_k(\sigma)$ convergences to 0. In order to avoid the possible singularity, we approximate the logarithmic function as

$$\log(T_{ij}) \approx \mathrm{PL}_\eta(T_{ij}) = \frac{1}{\eta}\left(T_{ij}^\eta - 1\right) \tag{22}$$

when calculating $CV_k(\sigma)$, where $\eta \in (0,1)$ is a small number. It is easy to prove that

$$\mathrm{PL}_\eta(x) - \log x = \frac{(\log x)^2}{2}\eta + O\left(\eta^2\right), \quad \text{for } x > 0 \tag{23}$$

and

$$\mathrm{PL}_\eta(x) \geq -\frac{1}{\eta}, \quad \text{for } x \geq 0 \tag{24}$$

Thus, the function $\mathrm{PL}_\eta(x)$ is bounded by $-1/\eta$ and approximately equal to $\log x$ if $x \in (0,1]$ is not close to 0.

4 Estimation with Stochastic Observations

In this section, we will take into account that the actual state transitions are unknown, and only stochastic observations

$$o_t|x_t \sim p\left(o_t|x_t\right) \tag{25}$$

for $t = 0, \ldots, m$ are available. In this case, the MC model of $\{x_t\}$ together with the above probability distributions describing the possible observations constitute a hidden Markov model (HMM) [32]. The MAP estimator of the TM with prior parameter σ can be expressed by

$$\begin{aligned} \hat{\boldsymbol{T}}\left(\sigma\right) &= \arg\max_{\boldsymbol{T}\in\Omega_n} \left\{\log p\left(O|\boldsymbol{T}\right) + \log p_{GMC}\left(\boldsymbol{T}|\sigma\right)\right\} \\ &= \arg\max_{\boldsymbol{T}\in\Omega_n} \left\{\log p\left(O|\boldsymbol{T}\right) - u\left(\boldsymbol{T},\sigma\right)\right\} \end{aligned} \tag{26}$$

where $O = \{o_0, \ldots, o_m\}$, and computed with the expectation maximization (EM) algorithm [33] consisting of the following steps:

Step 1. Choose an initial $\boldsymbol{T}^{(0)} \in \Omega_n$ and let $k = 0$.

Step 2. Compute the functional

$$\begin{aligned} Q\left(\boldsymbol{T}|\boldsymbol{T}^{(k)}\right) &= \mathbb{E}\left[\log\left(\boldsymbol{C}\left(X\right)|\boldsymbol{T}\right) - u\left(\boldsymbol{T},\sigma\right)|\boldsymbol{T}^{(k)}, O\right] \\ &= \sum_{i,j} \bar{C}_{ij} \log T_{ij} - u\left(\boldsymbol{T},\sigma\right) \end{aligned} \tag{27}$$

where $X = \{x_0, \ldots, x_m\}$, $\boldsymbol{C}\left(X\right) = \left[C_{ij}\left(X\right)\right]$ denotes the count matrix of X, and

$$\bar{\boldsymbol{C}} = \left[\bar{C}_{ij}\right] = \mathbb{E}\left[\boldsymbol{C}\left(X\right)|\boldsymbol{T}^{(k)}, O\right] \tag{28}$$

can be calculated by the forward-backward procedure [34].

Step 3. Find $\boldsymbol{T}^{(k+1)}$ which maximizes the function $Q\left(\boldsymbol{T}|\boldsymbol{T}^{(k)}\right)$ as

$$\begin{aligned} \boldsymbol{T}^{(k+1)} &= \arg\max_{\boldsymbol{T}\in\Omega_n} Q\left(\boldsymbol{T}|\boldsymbol{T}^{(k)}\right) \\ &= \arg\min_{\boldsymbol{T}\in\Omega_n} \left\{-\sum_{i,j} \bar{C}_{ij} \log T_{ij} + u\left(\boldsymbol{T},\sigma\right)\right\} \end{aligned} \tag{29}$$

Step 4. Terminate if

$$\left|\left(\log p\left(O|\boldsymbol{T}^{(k+1)}\right) - u\left(\boldsymbol{T}^{(k+1)},\sigma\right)\right) - \left(\log p\left(O|\boldsymbol{T}^{(k)}\right) - u\left(\boldsymbol{T}^{(k)},\sigma\right)\right)\right| \leq \epsilon_2 \tag{30}$$

where ϵ_2 is a small positive number.

Step 5. Let $k = k + 1$ and go to Step 2.

Note that (29) has the same form as (16) with $\bar{C}_{ij} \geq 0$ for any i, j, so (29) is a convex optimization problem and can be solved by the algorithm described in Appendix A too.

Further, in a similar manner to Section 3.3, the value of σ can be designed through the following MCCV algorithm:

Step 1. Divide O into M parts $O_1, \ldots, O_M$ with equal size as

$$O_i = \left\{o_t|\frac{\left(i-1\right)\left(m+1\right)}{M} \leq t < \frac{i\left(m+1\right)}{M}\right\} \tag{31}$$

Step 2. Partition $\{O_1, \ldots, O_M\}$ randomly into train subset S^{train} and test subset S^{test} with

$$p\left(O_i \in S^{train}\right) = \beta \tag{32}$$

for $i = 1, \ldots, M$.

Step 3. Calculate

$$\hat{\boldsymbol{T}}_k(\sigma) = \arg\max_{\boldsymbol{T}\in\Omega_n} \left\{\log p\left(S^{train}|\boldsymbol{T}\right) - u(\boldsymbol{T},\sigma)\right\} \tag{33}$$

by EM algorithm, and the predictive log-likelihood

$$CV_k(\sigma) = \log p\left(S^{test}|\hat{\boldsymbol{T}}_k(\sigma), S^{train}\right) = \log p\left(S^{test}, S^{train}|\hat{\boldsymbol{T}}_k(\sigma)\right) - \log p\left(S^{train}|\hat{\boldsymbol{T}}_k(\sigma)\right) \tag{34}$$

by the likelihood evaluation algorithm of HMMs [34].

Step 4. Repeat Steps 2 and 3 for $k = 1, \ldots, K$ and select

$$\sigma^* = \arg\max_{\sigma} CV(\sigma) \tag{35}$$

with $CV(\sigma) = \sum_k CV_k(\sigma)/K$.

Remark 2 *In this paper, we discuss the estimation of the TM, given a unique state sequence* $\{x_1, \ldots, x_m\}$ *or observation sequence* $\{o_1, \ldots, o_m\}$*. It is easy to extend our framework to the case of mutiple state or observation sequences. We only need to point out that for multiple independent stochastic observation sequences* $O^1, \ldots, O^M$*, we can treat each sequence* O^i *as a part when performing the MCCV of* σ*, i.e., revise* (31) *as* $O_i = O^i$*, and the predictive log-likelihood in* (34) *can be written as*

$$CV_k(\sigma) = \log p\left(S^{test}|\hat{\boldsymbol{T}}_k(\sigma), S^{train}\right) = \log p\left(S^{test}|\hat{\boldsymbol{T}}_k(\sigma)\right) \tag{36}$$

5 Simulations

5.1 Brownian Dynamics Model

In this section, the estimation method proposed in this paper will be applied to a Brownian dynamics (BD) model, which is described as

$$\mathrm{d}r = -f(r)\,\mathrm{d}t + \rho \mathrm{d}W \tag{37}$$

where $\rho = 1.4$, W is a standard Brownian motion, $f(r) = \mathrm{d}V(r)/\mathrm{d}r$ and $V(r)$ is the potential function (see Fig. 1) given by

$$V(r) = \begin{cases} -111.01r^3 + 178.63r^2 - 82.27r + 10.55, & r < 0.75 \\ 182.8915r^3 - 482.64r^2 + 413.69r - 113.44, & 0.75 \le r < 1 \\ -153.36r^3 + 526.11r^2 - 595.06r + 222.81 & 1 \le r < 1.25 \\ 84.94r^3 - 367.53r^2 + 521.98r - 242.62, & 1.25 < r \end{cases} \tag{38}$$

Using Euler-Maruyama scheme [35], the motion equation (37) can be discretized with time step $\Delta t = 10^{-3}$ as

$$r(t+\Delta t)\,|\,r(t) \sim \mathcal{N}\left(r(t) - \Delta t \cdot f(r(t)), \rho^2 \Delta t\right) \tag{39}$$

where $\mathcal{N}\left(\mu, v^2\right)$ denotes the normal distribution with mean μ and variance v^2. And the state space of r can be decomposed into $n = 100$ "cells" $S = \{s_1, \ldots, s_n\}$ with

$$x_k = s_i \Leftrightarrow \mathrm{sat}(r(k\Delta t), 0, 2) \in \left(\frac{2(i-1)}{n}, \frac{2i}{n}\right) \tag{40}$$

and

$$s_i = \frac{2i-1}{n} \tag{41}$$

where

$$\mathrm{sat}(r, l, u) = \begin{cases} l, & r < l \\ u, & r > u \\ r, & \text{otherwise} \end{cases} \tag{42}$$

Then the grid based approximate MC model can be expressed as

$$p(x_{k+1} = s_j | x_k = s_i) \propto \exp\left(-\frac{(s_j - s_i + \Delta t f(s_i))^2}{2\rho^2 \Delta t}\right) \tag{43}$$

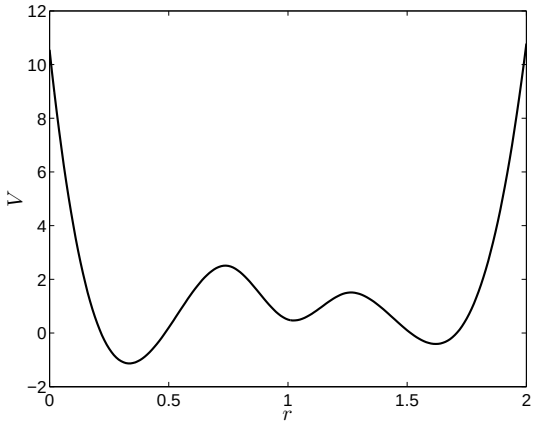


Figure 1: Potential Function

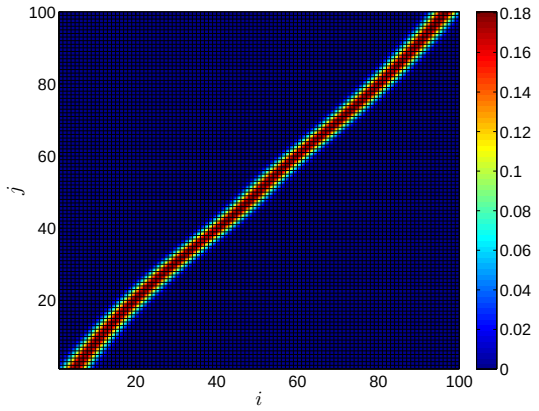


Figure 2: $\boldsymbol{T}$

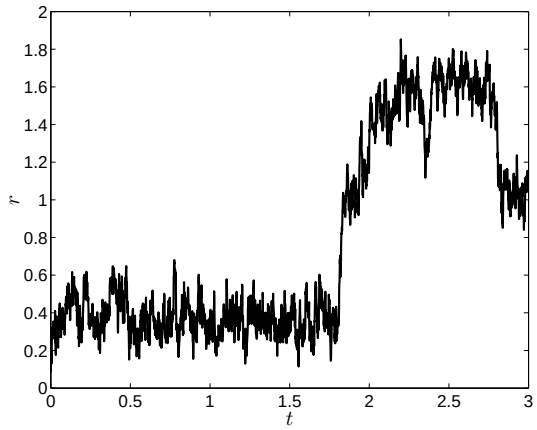


Figure 3: $r(t)$

The corresponding TM $\boldsymbol{T} = [T_{ij}]$ is shown in Fig. 2.

Furthermore, the neighborhood structure on S is here defined by

$$\partial s_i = \{s_{i-1}, s_{i+1}\} \cap S \tag{44}$$

with distance measure

$$d(s_i, s_j) = |i - j| \tag{45}$$

5.2 TM Estimation

We generate a realization $\{r(t) | 0 \leq t \leq 3\}$ of (37) with $r(0)$ randomly chosen from a uniform distribution over $[0, 2]$ (see Fig. 3), and discretize it into $\{x_k | 0 \leq k \leq m\}, m = 3000$ by using the discretization method in Subsection 5.1.

Here, we will use the MAP method presented in Section 3 to estimate the TM $\boldsymbol{T}$ based on $\{x_k | 0 \leq k \leq m\}$ and compare it with the ML method [15]. The algorithm parameters are chosen as

$$\epsilon_1 = 10^{-6}, \beta = 0.5, \eta = 0.1, K = 20 \tag{46}$$

and σ is selected from

$$S_\sigma(\sigma_l, \sigma_u) = \left\{ \exp\left(\frac{i}{19} \log \frac{\sigma_u}{\sigma_l} + \log \sigma_l \right) | i = 0, \ldots, 19 \right\} \subset [\sigma_l, \sigma_u] \tag{47}$$

with $\sigma_l = 10^{-2}$ and $\sigma_u = 10^{-0.5}$. Fig. 4 plots the MCCV results of σ and the optimal $\sigma^* = 0.06159$.

The comparisons of the different estimators are based on the Kullback-Leibler (KL) divergence rate

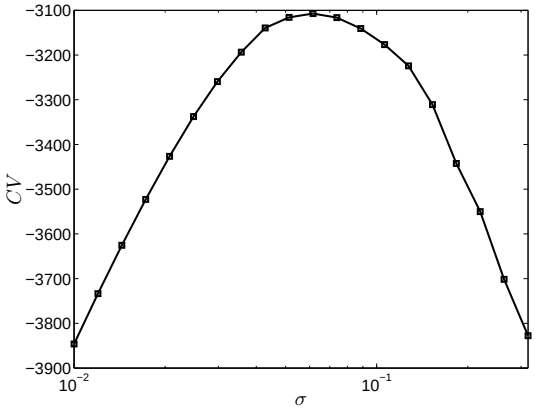


Figure 4: $CV(\sigma)$

Table 1: KL divergence rate metric between $\hat{\boldsymbol{T}}$ and $\boldsymbol{T}$

Estimator	$KLR\left(\hat{\boldsymbol{T}}\|\boldsymbol{T}\right)$
MAP ($\sigma = \sigma^*$)	0.0872
MAP ($\sigma = 3\sigma^*$)	0.1232
MAP ($\sigma = \sigma^*/3$)	0.7290
ML	0.2316

metric [36] defined as

$$\begin{aligned} KLR\left(\hat{\boldsymbol{T}}\|\boldsymbol{T}\right) &= \lim_{m\to\infty}\frac{1}{m}\sum_{x_0,\ldots,x_m} p\left(x_0,\ldots,x_m|\hat{\boldsymbol{T}}\right)\log\frac{p\left(x_0,\ldots,x_m|\hat{\boldsymbol{T}}\right)}{p\left(x_0,\ldots,x_m|\boldsymbol{T}\right)} \\ &= \sum_{ij}\hat{\pi}_i\hat{T}_{ij}\log\frac{\hat{T}_{ij}}{T_{ij}} \end{aligned} \tag{48}$$

where $\hat{\pi} = [\hat{\pi}_i]$ denotes the stationary distribution of TM $\hat{T} = \left[\hat{T}_{ij}\right]$. It can measure the distances between Markov chains on the same state space.

Fig. 5 shows the estimation results of the proposed MAP method with different σ and ML method, and Table 1 shows the estimation errors. Clearly, the ML method fails to estimate the values T_{ij} with $i \in [1,7]\cup[34,44]\cup[91,100]$ for there are few x_k are sampled within the ranges. The GMRF prior based MAP estimator overcome this problem by interpolating from the other T_{ij} according to the GMRF model. Moreover, as observed from the figures, the parameter σ determines the overall smoothness of the estimated TM, and the MCCV approach can provide an appropriate value of σ.

5.3 TM Estimation with Noisy Data

In this subsection, we study the performance of our proposed algorithms for estimating $\boldsymbol{T}$ from noisy observations

$$o(t)|r(t) \sim \mathcal{N}\left(r(t), v^2\right) \tag{49}$$

with $v = 0.1$. Fig. 6 shows a realization of $y(t)$. And the grid based approximate observation model is set as

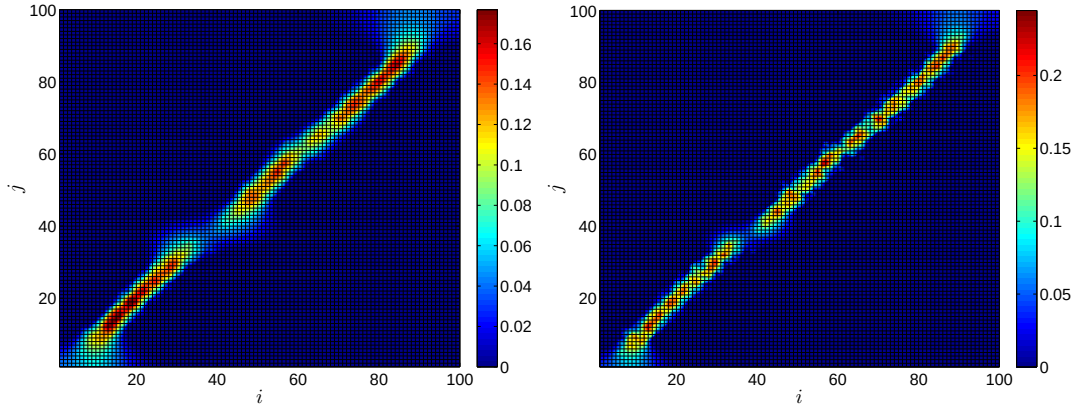
$$o_k|x_k = s_i \sim \mathcal{N}\left(s_i, v^2\right) \tag{50}$$

It is worth pointing out that the ML estimator

$$\hat{\boldsymbol{T}}^{ML} = \left[\hat{T}_{ij}^{ML}\right] = \arg\max_{\boldsymbol{T}\in\Omega_n}\log p\left(o_0,\ldots,o_m|\boldsymbol{T}\right) \tag{51}$$

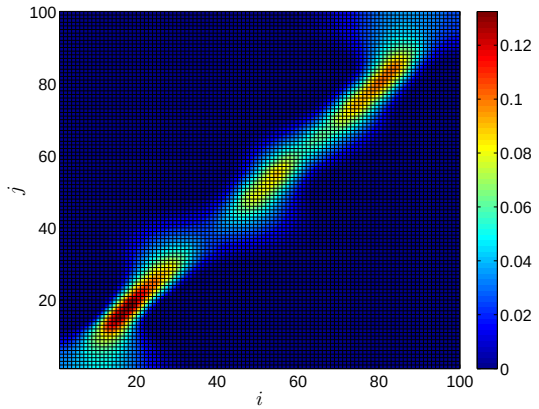
tends to over-fit the data for this estimation problem if $1/N$ is small enough. In the extreme case where

$$I_k = \arg\max_i p\left(o_k|s_i\right) \neq I_l = \arg\max_i p\left(o_l|s_i\right), \quad \forall k,l \in [0,m], k \neq l \tag{52}$$

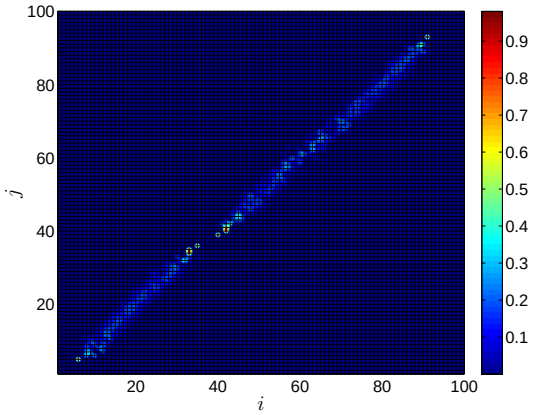


(a) MAP ($\sigma = \sigma^*$)

(b) MAP ($\sigma = 3\sigma^*$)



(c) MAP ($\sigma = \sigma^*/3$)



(d) ML

Figure 5: $\hat{\boldsymbol{T}}$

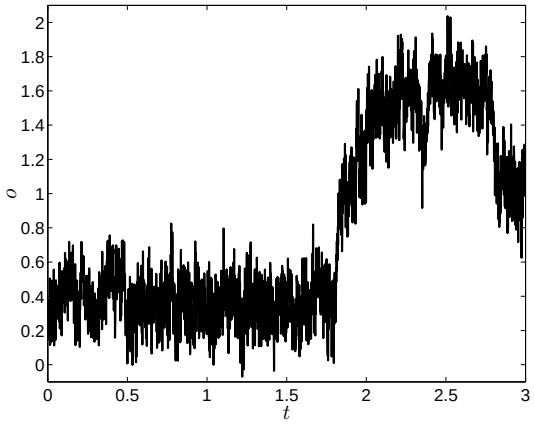


Figure 6: $o(t)$

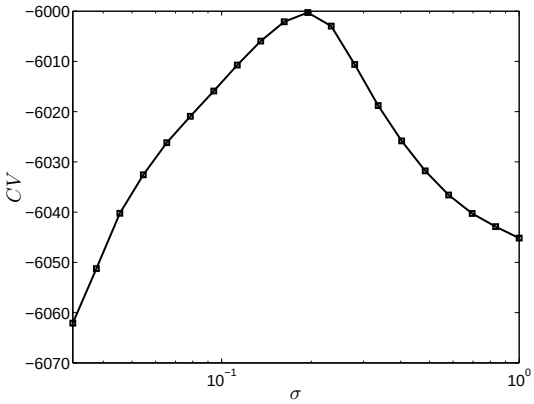


Figure 7: $CV(\sigma)$

it is easy to prove that the estimated $\hat{\boldsymbol{T}}^{ML}$ will be

$$\hat{T}_{ij}^{ML} = \begin{cases} 1, & I_k = i, I_{k+1} = j \text{ for some } k \\ 0, & \text{otherwise} \end{cases} \tag{53}$$

which implies that $\{x_0, \ldots, x_m\} = \{s_{I_0}, \ldots, s_{I_m}\}$ is a determinate process.

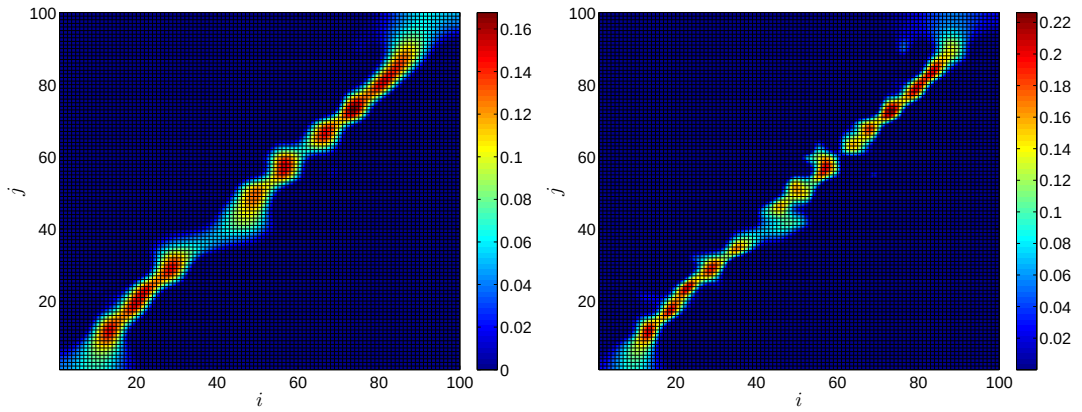
The MAP estimator with GMRF prior in 4 will now be compared to the ML estimator implemented using Baum-Welch algorithm [37]. The parameters of MAP estimator are set as

$$\epsilon_1 = 10^{-6}, \epsilon_2 = 10^{-2}, \beta = 0.5, K = 20, M = 10 \tag{54}$$

and σ is selected from $S_\sigma\left(10^{-1.5}, 1\right)$. The MCCV results are shown in Fig. 7 and the optimal $\sigma^* = 0.1947$.

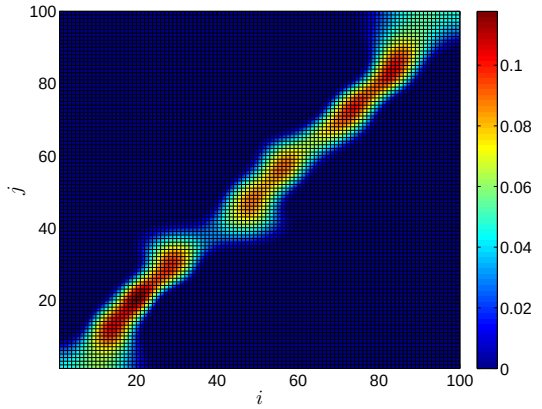
In Fig. 8, we show plots for $\hat{\boldsymbol{T}}$ obtained using our MAP estimator with $\sigma = \sigma^*$, $3\sigma^*$ and 0.06159 (σ^* in Subsection 5.2) and ML estimator. And the corresponding $KLR\left(\hat{\boldsymbol{T}} \| \boldsymbol{T}\right)$ are listed in Table 2. As can be seen from Fig. 8d, the ML estimator exhibits strong overfitting, and there are 58 states s_i of the total 100 states satisfying $\max_j \hat{\boldsymbol{T}}_{ij}^{ML} \geq 0.8$ (the maximum entry of $\boldsymbol{T}$ is only 0.1841). With comparison to ML method, the proposed MAP estimator avoids overfitting by adding a regularization term $u(\boldsymbol{T}, \sigma)$ (see (26) and (51)), which penalizes excessively large value of T_{ij}. And the best estimation performance is achieved when $\sigma = \sigma^*$.

Note that here $\sigma^* = 0.1947$ is bigger than the $\sigma^* = 0.06159$ in the previous subsection, which may be related to noisy observation and insufficient sample size. From Figs. 5a and 8c, we can see that the observation noise makes $\hat{\boldsymbol{T}}$ obtained from $\{o_0, \ldots, o_m\}$ smoother than that directly estimated by states $\{x_0, \ldots, x_m\}$. Therefore the MCCV approach will select a bigger σ^* to get a suitably smooth $\hat{\boldsymbol{T}}$ and maximize the predictive likelihood.

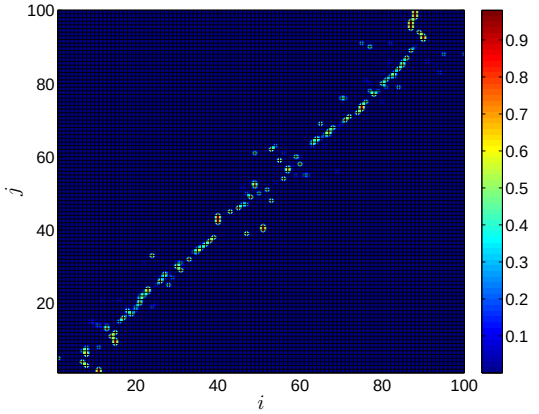


(a) MAP ($\sigma = \sigma^*$)

(b) MAP ($\sigma = 3\sigma^*$)



(c) MAP ($\sigma = 0.06159$)



(d) ML

Figure 8: $\hat{\boldsymbol{T}}$

Table 2: KL divergence rate metric between $\hat{\boldsymbol{T}}$ and $\boldsymbol{T}$

Estimator	$KLR\left(\hat{\boldsymbol{T}}\|\boldsymbol{T}\right)$
MAP ($\sigma = \sigma^*$)	0.1700
MAP ($\sigma = 3\sigma^*$)	0.2406
MAP ($\sigma = 0.06159$)	0.4902
ML	2.2819

6 Conclusions and Future Works

The GMRF model of TMs discussed in this paper provides a general and flexible framework for analyzing and estimating MCs with "smooth" TMs by extending the neighborhood relationship between states to that between transition pairs. This model is helpful to improve the robustness and accuracy of estimators in many practical cases, especially when the sample size is small with respect to the size of state space. And the convex form of GMRF model benefits the numerical calculation. The parameter choice is a difficult problem for our model, but it can be solved by CV methods since there is only one undetermined parameter. The models and methods presented in this paper can be modifed so that one can explore other extensions.

1. The Bayesian estimation based on the GMRF prior model. Theoretically speaking, the posterior distribution of TM can be approximated by numerical sampling methods. But the entries of TM are highly coupled under the assumption of GMRF prior, and commonly used Monte Carlo Markov Chain (MCMC) methods (e.g. Gibbs sampling) may be inefficient for this problem. Moreover, the full Bayesian estimation of TM and the parameter σ is also difficult for intractable $z(\sigma)$.
2. Application of other Markov random fileds (MRFs), such as generalized GMRF [38] which is more flexible than GMRF, and divergence potential function based MRF [39] which is more suitable to describe the variation of probability values.

These will be considered in future researches.

A Solver

Consider the following convex optimization

$$\begin{aligned} &\min_{\boldsymbol{T}} f(\boldsymbol{T}) \\ &\text{s.t.} \sum_j T_{ij} = 1, \quad i = 1, \ldots, n \\ &\quad T_{ij} \geq 0, \quad i, j = 1, \ldots, n \end{aligned} \tag{55}$$

In this paper, we solve the optimization problem as (55) using the diagonalized Newton (DN) method (see [40, 41] for details), which has a good practical rate of convergence, and takes advantage of the structure of the problem. The algorithm consists of the following steps:

Step 1. Choose an initial feasible solution $\boldsymbol{T}^{(0)}$ and let $k = 0$, $f_{-1}^L = -\infty$.

Step 2. Approximate the optimization problem (55) with the DN subproblem

$$\begin{aligned} &\min_{\boldsymbol{T}} \hat{f}(\boldsymbol{T}) = \left(\sum_{i,j} \frac{1}{2} a_{ij}^{(k)} \left(T_{ij} - T_{ij}^{(k)} \right)^2 + b_{ij}^{(k)} \left(T_{ij} - T_{ij}^{(k)} \right) \right) + f\left(\boldsymbol{T}^{(k)} \right) \\ &\text{s.t.} \sum_j T_{ij} = 1, \quad i = 1, \ldots, n \\ &\quad T_{ij} \geq 0, \quad i, j = 1, \ldots, n \end{aligned} \tag{56}$$

where

$$a_{ij}^{(k)} = \frac{\partial^2 f\left(\boldsymbol{T}^{(k)} \right)}{\partial T_{ij}^2}, \quad b_{ij}^{(k)} = \frac{\partial f\left(\boldsymbol{T}^{(k)} \right)}{\partial T_{ij}} \tag{57}$$

and compute $\boldsymbol{T}'$ as a solution to (56).

Step 3. Let $\boldsymbol{D} = \boldsymbol{T}' - \boldsymbol{T}$ and

$$f_k^L = \max \left\{ f_{k-1}^L, f\left(\boldsymbol{T}^{(k)} \right) + b_{ij}^{(k)} \left(T'_{ij} - T_{ij}^{(k)} \right) \right\} \tag{58}$$

Step 4. Apply Armijo rule [42] to obtain

$$\tau_k \approx \arg\max_{\tau\in[0,1]} f\left(\boldsymbol{T}^{(k)} + \tau \boldsymbol{D}\right) \tag{59}$$

and let $\boldsymbol{T}^{(k+1)} = \boldsymbol{T}^{(k)} + \tau_k \boldsymbol{D}$, $f_k^U = f\left(\boldsymbol{T}^{(k+1)}\right)$.

Step 5. Terminate if $\left|f_k^U - f_k^L\right| < \epsilon_1$, where ϵ_1 is a small positive numbers.

Step 6. Let $k = k + 1$ and go to Step 2.

For the theoretically optimal solution $\boldsymbol{T}^*$, it can be proved that f_k^L and f_k^U are lower and upper bounds of $f\left(\boldsymbol{T}^*\right)$, and the gap $\left|f_k^U - f_k^L\right|$ convergences to 0 as $k \to \infty$. Moreover, the DN subproblem (56) can separate into n small-scale quadratic programming problems for $i = 1, \ldots, n$ as

$$\begin{aligned} \min_{T_{i1},\ldots,T_{in}} & \sum_{j=1}^{n} \frac{1}{2} a_{ij}^{(k)} \left(T_{ij} - T_{ij}^{(k)}\right)^2 + b_{ij}^{(k)} \left(T_{ij} - T_{ij}^{(k)}\right) \\ \text{s.t.} & \sum_{j=1}^{n} T_{ij} = 1 \\ & T_{ij} \geq 0, \quad j = 1, \ldots, n \end{aligned} \tag{60}$$

And each of them can be solved analytically via its Lagrangian dual problem.

References

[1] G. Bowman, K. Beauchamp, G. Boxer, and V. Pande, "Progress and challenges in the automated construction of Markov state models for full protein systems," *The Journal of Chemical Physics*, vol. 131, p. 124101, 2009.

[2] A. Ezzati, "Forecasting market shares of alternative home-heating units by Markov process using transition probabilities estimated from aggregate time series data," *Management Science*, vol. 21, no. 4, pp. 462–473, 1974.

[3] B. Craig and P. Sendi, "Estimation of the transition matrix of a discrete-time Markov chain," *Health economics*, vol. 11, no. 1, pp. 33–42, 2002.

[4] J. Zhu, J. Hong, and J. Hughes, "Using Markov models for web site link prediction," in *Proceedings of the thirteenth ACM conference on Hypertext and hypermedia*. ACM, 2002, p. 170.

[5] L. Florence, "Skill Evaluation in Women's Volleyball," *Journal of Quantitative Analysis in Sports*, vol. 4, no. 2, 2008.

[6] G. Miller, "Finite Markov processes in psychology," *Psychometrika*, vol. 17, no. 2, pp. 149–167, 1952.

[7] A. Madansky, "Least squares estimation in finite Markov processes," *Psychometrika*, vol. 24, no. 2, pp. 137–144, 1959.

[8] L. Telser, "Least-squares estimates of transition probabilities," *Measurement in Economics*, pp. 270–292, 1963.

[9] T. Lee, G. Judge, and T. Takayama, "On estimating the transition probabilities of a Markov process," *Journal of Farm Economics*, vol. 47, no. 3, pp. 742–762, 1965.

[10] H. Theil and G. Rey, "A quadratic programming approach to the estimation of transition probabilities," *Management Science*, vol. 12, no. 9, pp. 714–721, 1966.

[11] G. Judge and T. Takayama, "Inequality restrictions in regression analysis," *Journal of the American Statistical Association*, vol. 61, no. 313, pp. 166–181, 1966.

[12] T. Anderson and L. Goodman, "Statistical inference about Markov chains," *The Annals of Mathematical Statistics*, vol. 28, no. 1, pp. 89–110, 1957.

[13] M. Kendall and A. Stuart, *The advanced theory of statistics*. London: Charles Griffin, 1961, vol. 2.

[14] T. Lee, G. Judge, and A. Zellner, *Estimating the parameters of the Markov probability model from aggregate time series data*. North-Holland, 1970.

[15] ——, "Maximum likelihood and Bayesian estimation of transition probabilities," *Journal of the American Statistical Association*, vol. 63, no. 324, pp. 1162–1179, 1968.

[16] J. Martin, *Bayesian decision problems and Markov chains*. Wiley New York, 1967.

[17] F. Noé, "Probability distributions of molecular observables computed from Markov models," *The Journal of Chemical Physics*, vol. 128, p. 244103, 2008.

[18] C. Fuh and T. Fan, "A Bayesian bootstrap for finite state Markov chains," *Statistica Sinica*, vol. 7, pp. 1005–1020, 1997.

[19] M. Meshkani and L. Billard, "Empirical Bayes estimators for a finite Markov chain," *Biometrika*, vol. 79, no. 1, pp. 185–193, 1992.

[20] S. Assoudou and B. Essebbar, "A Bayesian Model for Markov Chains via Jeffrey's Prior," *Communications in Statistics*, vol. 32, no. 11, 2003.

[21] ——, "A Bayesian model for binary Markov chains," *International Journal of Mathematics and Mathematical Sciences*, vol. 2004, no. 8, pp. 421–429, 2004.

[22] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 192–236, 1974.

[23] J. Besag and C. Kooperberg, "On conditional and intrinsic autoregressions," *Biometrika*, vol. 82, no. 4, pp. 733–746, 1995.

[24] H. Rue and L. Held, *Gaussian Markov random fields: theory and applications*. Chapman & Hall, 2005.

[25] S. Li, *Markov random field modeling in image analysis*. Springer, 2009.

[26] S. Saquib, C. Bouman, and K. Sauer, "ML parameter estimation for Markov random fields with applications to Bayesian tomography," *IEEE Transactions on Image Processing*, vol. 7, no. 7, pp. 1029–1044, 1998.

[27] J. Shao, "Linear model selection by cross-validation," *Journal of the American Statistical Association*, pp. 486–494, 1993.

[28] M. van der Laan, S. Dudoit, and S. Keles, "Asymptotic optimality of likelihood-based cross-validation," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, p. 1036, 2004.

[29] P. Smyth, "Clustering using Monte Carlo cross-validation," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 126–133.

[30] ——, "Clustering sequences with hidden Markov models," *Advances in Neural Information Processing Systems*, 1997.

[31] Q. Xu and Y. Liang, "Monte Carlo cross validation," *Chemometrics and Intelligent Laboratory Systems*, vol. 56, no. 1, pp. 1–11, 2001.

[32] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, pp. 14–16, 1986.

[33] A. Dempster, N. Laird, D. Rubin, *et al.*, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[34] R. Lawrence and A. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[35] P. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations.* Springer, 1995.

[36] Z. Rached, F. Alajaji, and L. Campbell, "The Kullback-Leibler divergence rate between Markov sources," *IEEE Transactions on Information Theory*, vol. 50, no. 5, pp. 917–921, 2004.

[37] L. Welch, "Hidden Markov models and the Baum-Welch algorithm," *IEEE Information Theory Society Newsletter*, vol. 53, no. 4, pp. 1–10, 2003.

[38] C. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Transactions on Image Processing*, vol. 2, no. 3, pp. 296–310, 1993.

[39] J. O'Sullivan, "Divergence penalty for image regularization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1994.

[40] T. Larsson, M. Patriksson, and C. Rydergren, "An efficient solution method for the stochastic transportation problem," in *Optimization Methods for Analysis of Transportation Networks, Linköping Studies in Science and Technology.* Department of Mathematics, Linköping University, 1998, no. 702.

[41] M. Daneva, T. Larsson, M. Patriksson, and C. Rydergren, "A comparison of feasible direction methods for the stochastic transportation problem," *Computational Optimization and Applications*, 2008.

[42] D. Bertsekas, *Nonlinear Optimization.* Belmont: Athena Scientific, 1999.